Disentangling the effects of attentional weighting and associative mediation in perceptual learning reveals no evidence for associative mediation

Abbreviated title: Attention and mediation in perceptual learning

David N. George [a], Bianca P. Oltean [a, b]

[a] Department of Psychology, University of Hull, Hull, UK

[b] Department of Psychiatry, University of Cambridge, Cambridge, UK

Address for correspondence:
David N. George
Department of Psychology
University of Hull
Hull HU6 7RX
Tel: +44 (0) 1482 465483
Email: d.george@hull.ac.uk

Keywords: attentional weighting, associative mediation, category learning, perceptual learning, similarity

Abstract

Learning to categorize perceptually similar stimuli can result in people becoming more sensitive to differences along perceptual dimensions that are relevant to category membership and/or less sensitive to equivalent differences along irrelevant perceptual dimensions. These effects of acquired distinctiveness and acquired equivalence may be caused by changes in the representations of stimuli which come about through adjustment to the relative attentional weighting of perceptual features or dimensions. Alternatively, the development of associations between individual stimuli and category labels could result in those labels being incorporated into the stimulus representation, hence increasing or decreasing generalization between the stimuli. For many categorization tasks, the expected effects of attentional weighting and associative mediation on stimulus similarity are the same. We report three experiments using complex category structures, which allowed us to assess the independent influence of each mechanism on stimulus similarity. The results suggest that, in these categorization tasks, attentional weighting affects perceptual similarity but associative mediation does not.

Disentangling the effects of attentional weighting and associative mediation in perceptual learning reveals no evidence for associative mediation

Experience can affect our ability to differentiate between stimuli, and our judgements of their similarity. Such perceptual learning effects are observed across many domains. Commonly cited examples include wine (Bende & Nordin, 1997; Walk, 1966) and beer tasting (Peron & Allen, 1988), the detection of features in X-ray images (Sowden, Davies & Roling, 2000), sexing of day old chicks (Biederman & Shiffrar, 1987), the identification of faces (Shapiro & Penrod, 1986), and differentiation between speech sounds (Liberman, Harris, Hoffman & Griffith, 1957). Typical findings are that category learning can increase the similarity of stimuli belonging to the same category (acquired equivalence; e.g., Livingstone, Andrews & Harnad, 1998), and/or decrease the similarity of stimuli belonging to different categories (acquired distinctiveness; e.g., de Leeuw, Andrews, Livingstone & Chin, 2016; Notman, Sowden & Özgen, 2005). For example, Goldstone (1994) trained participants to sort squares differing in size and brightness into two categories based on one or other of those perceptual dimensions. In a subsequent same-different judgement task, these participants were more sensitive to a difference on the category relevant dimension than control participants who had not completed the categorization task. This effect was especially large at the boundary between the categories.

An early explanation of acquired distinctiveness was put forward by James (1890), suggesting that the discriminability of two very similar stimuli might be enhanced if they are associated with other events which are quite different from each other. As an illustrative example, he considered how one could come to distinguish between claret and burgundy, which might at first encounter taste very similar. He reasoned that the wines will be

consumed on different occasions and referred to by their different names. Consequently, the taste of each will come to recall its name, and memories of the occasions on which we have previously drunk that wine. Eventually, as they are consumed in many different environments, the importance of these incidental memories will fade. At the same time, however, the association of the taste of each wine with the distinct names will become inveterate. This is what solidifies the discrimination between the two tastes. 'The names differ far more than the flavors, and help to stretch these latter farther apart' (James, 1890, p. 511). Building on the principle of secondary generalization (Hull, 1939), Miller and Dollard (1941) developed a related account of how different stimuli associated with the same response might acquire equivalence, and those associated with different responses acquire distinctiveness (Miller, 1948).

Similar processes of associative learning formed the basis of an account of acquired equivalence and distinctiveness described by Hall (1991). A stimulus representation may be thought of as consisting of a collection of features (e.g., Estes, 1950; Konorski, 1948), with the representations of similar stimuli sharing some features in common. Hence, two stimuli, A and B, may be represented by three sets of features: those unique to A, those unique to B, and those common to A and B. If A and B are both paired with the same stimulus or event, X, then subsequent presentation of either stimulus will associatively activate the representation of X. Effectively, the proportion of features that stimuli A and B have in common is increased by their common association with X and they will have acquired equivalence. Similarly, if A and C are associated with different stimuli or events, X and Y, the representations of which share fewer common features than those of A and C do, the perceived overlap between stimuli A and C may be reduced and they will acquire distinctiveness. Central to this proposal,

in common with those of James (1890) and Miller and Dollard (1941; Miller, 1948), is the notion that the similarity of stimuli, or generalization between them, is affected (or mediated) by their associations with other events. In effect, the associates of stimuli become incorporated into the stimulus representations, thereby either increasing or decreasing the amount of overlap between those representations. Hence, we may refer to these mechanisms as associative mediation.

Hall, Mitchell, Graham and Lavis (2003), found evidence that the associates of stimuli can indeed affect subsequent generalization between them. They trained participants to associate four geometric shapes (A, B, C, and D) with two coloured rectangles (X, and Y). Shapes A and B were each always followed by rectangle X, and shapes C and D were each followed by rectangle Y. In a second phase of the experiment, one group of participants (Group Consistent) were trained to make a left key press in response to A or B, and a right key press in response to C or D. Hence, the pairs of shapes that were followed by a particular rectangle in the first phase of the experiment required the same response in the second phase. Participants in Group Inconsistent received similar training, but they had to press the left key in response to A or C, and the right key in response to B or D. For these participants, pairs of shapes that were followed by the same rectangle in the first phase of training required different responses in the second phase. In a final test, participants in both groups were given the opportunity to make left or right key presses in response to rectangles X and Y. Participants in Group Consistent acquired the second phase task more rapidly than those in Group Inconsistent and, on the test trials, tended to press the left key in the presence of stimulus X and the right key in the presence of Y. Participants in Group Inconsistent were equally likely to press either key in the presence of each rectangle. This pattern of results

suggest that, for Group Consistent, a mental representation of rectangle X was retrieved when either shape A or B was presented during the second phase and this representation became associated with the left key press response. Similarly, an associatively activated representation of rectangle Y became associated with the right key press on trials on which shape C or D was presented.

Associative mediation may also explain the effect of verbal labels on category learning. Lupyan, Rakison & McClelland (2007) trained participants to make approach or avoid responses to members of two categories of complex three-dimensional shapes which differed from each other in several aspects. A group for which redundant verbal category labels were presented along with each stimulus showed faster learning and better categorization performance on test trials even when the labels were omitted. These results suggest that the stimuli may activate the label with which they are associated and the label, in turn, exerts a top-down effect on perceptual representations (Lupyan, 2012). Consistent with this suggestion, Russian speakers perform better than English speakers in a speeded colour discrimination between blues which have different names in Russian, but do not in English. This advantage disappears when activation of the colours' labels is inhibited by a concurrent verbal (but not spatial) dual-task (Winawer et al, 2007).

Attentional weighting[1] provides an alternative explanation of acquired equivalence and distinctiveness effects. Lawrence (1950; see also 1949) trained rats in an apparatus

[1] The concept of attention is frequently invoked in cognitive psychology without a great deal of explanation of what is meant by the term and attempts are seldom made to measure attention directly (see Quinlan, 2010 for a discussion of the confusion surrounding the word in one context). Here, we use the term attentional weighting as a synonym for dimensional weighting for consistency with the existing literature (e.g., Goldstone, 1998).

consisting of two parallel runways that were either both black or both white and either contained curtains of chains or did not. The rats were rewarded for making particular responses based on one of these perceptual dimensions. For example, choosing the left runway when they were both black, and the right runway when they were both white. The second dimension was uncorrelated with reward. In a subsequent phase of the experiment, rats learned to discriminate between the previously relevant stimuli more readily than the previously irrelevant stimuli. Lawrence (1952) suggested that the importance of this type of discrimination learning is that it allows an animal to isolate the relevant stimulus dimension from other irrelevant cues.

This latter idea formed the basis of Gibson and Gibson's (1955; Gibson, 1969) account of perceptual learning. They argued that the associations formed during discrimination learning are not responsible for acquired equivalence and distinctiveness effects (as suggested by the associative mediation account), but rather discrimination training makes the subject attend to differences between stimuli which were previously undetected, and to ignore irrelevant features. Hence, perceptual learning involves a change in emphasis of different aspects of the stimulus, rather than a change in the content of the representation. Sutherland and Mackintosh (1971; see also Zeaman & House, 1963; Lovejoy, 1968) described a formal model of attention in discrimination learning, and similar processes of selective attention are implemented in models of categorization such as GCM (Nosofsky, 1986) and ALCOVE (Kruschke, 1992) which propose that psychological space may be stretched along stimulus dimensions that are relevant to a categorization task and compressed along irrelevant dimensions.

Using an experimental design related to Hall et al's (2003), Bonardi, Graham, Hall and Mitchell (2005) found that attentional weighting may also affect acquired equivalence and distinctiveness in human learning. Two groups of participants received initial training in which four visual cues, two shapes and two colours, were each associated with one of two verbal labels. For participants in Group S, the two colours were associated with one label and the two shapes with the other. For Group D, one colour and one shape were associated with the first label and the other shape and colour with the second label. In a second phase of the experiment participants were then required to make either a left or a right key press in response to each stimulus. For all participants, one shape and one colour were paired with each response. The task was structured such that two stimuli associated with the same label in the first phase were paired with different responses in the second phase. This meant that associative mediation should predict no difference in the performance of the two groups. Group D learned the second task faster than Group S. Bonardi et al argued that the first phase of training encouraged participants in Group D to attend to the features that distinguished the two stimuli within each perceptual dimension (shape and colour) and that enhanced their ability to acquire different responses to them in the second phase.

Hall et al (2003) and Bonardi et al (2005) provided evidence in support of either associative mediation or attentional weighting using carefully controlled experimental designs which allow the influence of each mechanism to be isolated. These experiments, however, employed simple stimuli and transfer tests where the effects of initial training were observed on the acquisition of a subsequent discrimination. The majority of recent studies of acquired equivalence and distinctiveness in humans have used more complex stimuli and assessed the effects using measures such as similarity ratings, same-different judgements, or

an X-AB (or match-to-sample) task (e.g., de Leeuw et al, 2016; Goldstone, 1994; Goldstone & Steyvers, 2001; Livingstone et al, 1998; Notman et al, 2005). In these experiments, participants are typically trained to categorize stimuli that differ along two dimensions where the category boundary lies at the middle of one of the dimensions while the other dimension is irrelevant. Similarity of pairs of stimuli is assessed following category training and compared to participants who have either not undergone this training, or who have been trained on an unrelated categorization task. Acquired distinctiveness may manifest as increased discriminability across the whole of the relevant stimulus dimension, or just at the category boundary (expansion). Acquired equivalence might be observed as an increase in similarity along the irrelevant dimension and/or as an increase in similarity between pairs of stimuli that belong to the same category but differ on the relevant dimension (compression).

Both associative mediation and attentional weighting predict an increase in discriminability between categories, and a reduction in discriminability along the irrelevant stimulus dimension. Other predictions, however, are less clear. While some models of selective attention apply changes to entire stimulus dimensions (e.g., Kruschke, 1992; Nosofsky, 1986; Sutherland & Mackintosh, 1971), others restrict changes to individual stimuli (e.g., Kruschke, 2001a, 2001b; Mackintosh, 1975). Hence, attentional weighting is consistent with either an increase in discriminability along the entirety of the relevant dimension, or an increase in similarity between stimuli that differ on that dimension but belong to the same category. The latter of these effects is also predicted by associative mediation. Predictions are further complicated because exposure to the test stimuli (or to stimuli very similar to them) during category training might have effects on performance which are not directly related to category learning (e.g., unsupervised learning processes such as imprinting; see Goldstone,

2003). Finally, while experiments usually show that participants are more sensitive to a difference between members of different categories than those belonging to the same category, the pattern of effects responsible for this difference varies. These factors make it difficult to determine whether associative mediation or attentional weighting are responsible for a particular pattern of results, or, if both mechanisms operate in tandem, which effect is greater.

The purpose of the experiments reported here was to isolate the effects of associative mediation and attention weighting using a design similar to that employed by Goldstone (1994) and Goldstone and Steyvers (2001), but training participants on a more complex, conditional categorization task. In Experiment 1 the category structure was of the sort identified by Shepard, Hovland and Jenkins (1961) as a type II classification, shown in Figure 1a. Stimuli differed on three dimensions, two of which were relevant and the third of which was irrelevant. Kruschke (1996; see also George & Pearce, 1999) has previously demonstrated acquired distinctiveness effects following training on this type of category structure, but using a two-stage transfer design. Similarly, Mathy, Haldjian, Laurent and Goldstone (2013) have shown that changing stimulus salience in ways that mimic the expected effects of attentional weighting – increasing the salience of relevant stimulus dimensions, and decreasing the salience of the irrelevant stimulus dimension – facilitate type II categorization performance. Experiment 2 employed a variation of this category structure to test the generality of the findings from Experiment 1, and in Experiment 3 a different structure (type IV in Shepard et al's nomenclature) was used to isolate the effect of associative mediation.

<Figure 1 about here>

Experiment 1

The structure of the categorization task employed in Experiment 1 is shown in Figure 1a. Participants were presented with patterns that differed along three binary dimensions: [A, B], [P, Q], and [X, Y]. The first two of these dimensions were relevant to the solution of the categorization task, whereas the third dimension was irrelevant. Hence, patterns APx, APy, BQx, and BQy (represented by blue circles at the vertices of the cube in Figure 1a) belonged to Category 1, whereas patterns AQx, AQy, BPx, and BPy (represented by white circles) belonged to Category 2 (where upper case letters indicate the stimulus features belonged to a relevant dimension and lower case letters indicate that they belong to an irrelevant dimension). Figure 2 shows changes in the distances between stimuli which might result from mechanisms of (a) selective attention, and (b) associative mediation following category learning. Changes in attentional weighting should make participants more sensitive to differences on relevant dimensions, stretching them apart, and less sensitive to differences on the irrelevant dimension, compressing it. Associative mediation accounts of perceptual learning, however, make no such predictions about the stretching or compression of perceptual dimensions. Rather, associative mediation should make participants more sensitive to the difference between patterns when they belonged to different categories than when they belonged to the same category. In essence, this introduces a fourth dimension upon which stimulus similarity varies which corresponds to category membership. Test trials were given following the category learning phase of the experiment in order to determine which, if either, of these effects had the greatest magnitude.

<Figure 2 about here>

Participants' sensitivity to differences between patterns was assessed in a simple same-different judgement task. On each trial, two patterns were presented very briefly, one after the other, with a random masking display separating them. On some occasions the two patterns were the same. On other occasions they differed in their value on either one or two dimensions. For example, as seen in Figure 1a, pattern APx might be compared with itself (S [same]; APx), one of the patterns to which it differed on one relevant dimension (R; AQx, BPx), the pattern that differed only on the irrelevant dimension (I; APy), the pattern that differed on both relevant dimensions (RR; BQx), or one of the patterns to which it differed on the irrelevant dimension and one of the two relevant dimensions (IR; AQy, BPy). When the two patterns differed on a single dimension, attentional weighting and associative mediation make the same prediction: participants should be more sensitive to the difference between APx and a pattern to which it differs on a relevant dimension (R test trials) than to the difference between APx and a pattern to which it differs on the irrelevant dimension (I test trials). Not only should attentional weighting amplify any difference on a relevant dimension relative to the irrelevant dimension, but also any two patterns that differed on a single relevant dimension belonged to different categories whereas any two patterns that differed only on the irrelevant dimension belonged to the same category.

The critical test trials were those on which the patterns differed on two dimensions. When both of those dimensions were relevant to the categorization task (RR test trials; e.g., APx vs. BQx), attentional weighting should make them easier to tell apart than when one of the dimensions was irrelevant (IR test trials; e.g., APx vs. AQy, or APx vs. BPy). Associative mediation should have the opposite effect. This is because patterns that differed on two

relevant dimensions belonged to the same category, whereas those that differed on a single relevant dimension (plus the irrelevant dimension) belonged to different categories.

We may, therefore, consider any discrepancy in sensitivity to I and R differences to be the sum of the effects of these two mechanisms, whereas any discrepancy in sensitivity to IR and RR differences should reflect the difference between the two mechanisms. These relationships are reflected in Equations 1a and 1b where *I, R, IR* and *RR* refer to measures of sensitivity to those comparisons and *attention* and *mediation* refer to the effects of those mechanisms:

$$R - I = attention + mediation \quad (1a)$$

$$RR - IR = attention - mediation \quad (1b)$$

Equations 1a and 1b may be combined and rewritten as Equations 2a and 2b to allow us to derive estimates of the independent contribution of each mechanism to any perceptual learning effects observed at test.

$$attention = \frac{(R - I) + (RR - IR)}{2} \quad (2a)$$

$$mediation = \frac{(R - I) - (RR - IR)}{2} \quad (2b)$$

Method

Participants. A total of 40 undergraduate psychology students at the University of Hull served as participants. Before the start of the experiment they were randomly assigned in approximately equal numbers to three counterbalancing groups which differed in which stimulus dimension was irrelevant for the category learning task. Participants were compensated with either course credit (19) or a payment of £8 (21). Half of the participants

failed to meet the learning criterion of 75% correct responses over the final 64 trials of the categorization task. All data for these participants were excluded from further analyses. Of the remaining 20 participants, 8, 6, and 6 participants belonged to the counterbalancing conditions in which brightness of the square, horizontal line position, and vertical line position, served as the irrelevant dimension in the categorization task, respectively. The ages of these 20 participants ranged from 19 to 25 years (*M* = 20.2), 14 were female, and 17 were right handed.

Stimuli and Apparatus. The experiment was run using personal computers running the Windows 7 operating system (Microsoft Corporation, Redmond WA), and programmed in MATLAB (release 2014a; Mathworks Inc., Natick MA) using the Psychophysics Toolbox Version 3 (Brainard, 1997; Pelli, 1997; Kleiner, Brainard & Pelli, 2007). Stimuli were presented on 5:4 aspect ratio iiyama ProLite 1906S 48-cm TFT monitors (iiyama Corporation, Tokyo, Japan) with a native resolution of 1280 x 1024 pixels (w x h) and 60-Hz refresh rate connected to Nvidia GeForce GT 430 graphics cards (Nvidia Corporation, Santa Clara CA). Responses were made on a standard UK 105-key PC keyboard.

Each stimulus consisted of a grey square measuring 120 pixels on each side with a white bar measuring 15 x 240 pixels (w x h) superimposed upon it. The stimuli were displayed on a black screen, and subtended a viewing angle of 2.7° x 5.4° (w x h) from a viewing distance of 75 cm. Individual stimuli were created by performing transformations on a single base pattern. In the base pattern, shown in Figure 3a, the grey square had a brightness of 127 units on a scale of 0 (black) to 255 (white). The centre of the white bar was offset from the centre of the grey square by 12 pixels to the right and down. The base pattern could be transformed by varying values on three dimensions: the brightness of the square, the horizontal position

of the bar, and the vertical position of the bar. In the first phase of the experiment, patterns were created in pairs by applying symmetrical transformations to a single dimension of between 4 and 40 units (units of brightness or pixels) in each direction. Once a participant's detection threshold for differences on each dimension was separately determined, these values were used to create eight patterns by factorial combination of the two possible values on each of the three dimensions. Those eight patterns were used in the two subsequent phases of the experiment.

<Figure 3 about here>

In the first and third phases of the experiment, masking displays were randomly generated on each trial. These masks measured 600 pixels along each side. They were generated by drawing 500 grey squares and 500 white lines at random locations within the mask. The squares had sides of 120 pixels in length and the brightness of each square was randomly determined within the range of 87 to 167 units (127 ± 40). The white bars all had a brightness of 255 units and measured 15 x 240 pixels.

Procedure

Threshold estimation. Detection thresholds for differences on each of the three dimensions upon which the patterns could differ were estimated using a staircase procedure. Each dimension was tested in a separate block of trials, and the order in which these blocks were presented was determined randomly for each participant. For each dimension, three staircases were run concurrently in a randomly interleaved manner, with each staircase starting at a different value (4, 20, or 40 units).

On each trial, a pair of stimuli was generated by transforming the value of one of the three dimensions symmetrically by the current value of the staircase. For example, if the current value was 20 units and the dimension being manipulated was brightness, then the two patterns would consist of grey squares with brightness of 107 and 147 units (127 ± 20), but with the white bar in the same location in each pattern. The sequence of events on each trial is shown in Figure 3b. At the beginning of the trial, a white fixation cross was displayed in the centre of the screen for 1000ms. One of the patterns was then presented for 150ms on the left-hand side of the screen in a random location within an imaginary rectangle measuring 200 x 400 pixels (w x h), which was vertically centred on the screen and the right-hand side of which was 50 pixels to the left of the horizontal centre of the screen. A masking pattern was then displayed for 250ms after which the second pattern was presented on the right-hand side of the screen, in a random location within a second imaginary rectangle measuring 200 x 400 pixels. This rectangle was also centred vertically on the screen, but its left-hand edge was 50 pixels to the right of the horizontal centre of the screen. The screen then remained blank until the participant responded by pressing either the ‘s’ key if they believed that the two patterns were the same, or the ‘d’ key if they believed them to be different. Once a response had been made, the next trial began immediately with the display of the fixation cross. Which pattern was presented first on a trial was determined randomly so that it was not the case that, for example, patterns containing light squares always appeared on the left of the screen and those containing dark squared always appeared on the right of the screen. Responses made while either of the patterns or the mask was displayed were ignored.

When the participant made a correct responses (always by identifying the two patterns as being different), the value of the current staircase was sometimes reduced to

make the subsequent judgement involving that staircase more difficult. When an incorrect response was made, the value was always increased to make the next trial within that staircase easier. On occasion, one or more changes in stimulus value in one direction (e.g., increases) would be followed by a change in the opposite direction (i.e., a decrease). These points were termed 'reversals'. Training on a given staircase continued until 12 such reversals had been performed on that staircase. Once 12 reversals had been completed, that staircase was terminated, and testing continued on the remaining staircase(s). Once all three staircases terminated, the block of trials ended, and the participant was given the opportunity for a brief rest before the next block of trials commenced.

Prior to the first reversal, the value of the staircase was adjusted following a single correct or incorrect response. Following the first reversal, two successive correct responses were required before the value of the staircase was decreased. Hence, we implemented a one-up, two-down staircase method. The size of the increments and decrements to the staircase value were equal, but were dependent upon the number of reversals that had been made. Initially, changes were of four units. Following the second reversal this was reduced to two units and following the sixth reversal they were reduced again to one unit. This procedure allowed convergence upon a point at which the participant performed with an accuracy of approximately 65% correct judgements.

Following the end of the third block of trials, detection thresholds for each dimension were estimated. First, the average value of each staircase over the last five reversals of that staircase was calculated. The two staircases for each dimension that had average values closest to each other were then identified and the average value of those two staircases was found. It was these values that were used to create patterns for the next two phases of the

experiment. This procedure was automated so that the participants did not experience a break in the experimental session.

Category learning. Eight patterns were created by combining two values on each of the three dimensions upon which the stimuli could vary. These eight patterns were then segregated into two categories. The category structure is shown in Figure 1a. Two of the three dimensions were relevant to the solution of the categorization task whereas the third dimension was irrelevant. The identity of the irrelevant dimension was different for each of the three counterbalancing groups.

Each trial began with the presentation of a white fixation cross in the centre of the screen. After 750ms, the fixation cross was replaced with one of the eight patterns, also in the centre of the screen. The caption 'Does this pattern belong to category G or H?' was displayed above the pattern in 36pt white text. The caption remained on the screen until the participant responded by pressing either the 'g' or 'h' key. Following a correct response, the caption was replaced with the word 'Correct' in 60pt green text. Following an incorrect response, the caption was replaced with the word 'Incorrect' in 60pt red text. In either case, the correct answer was also presented; the words 'The pattern belongs to category [G/H].' were shown in 36pt white text below the pattern. This feedback remained on the screen for 2000ms after which the screen was cleared. There was then a 750ms interval before the next trial was initiated.

Within each block of eight trials, each of the eight patterns was presented once in a random order. Training continued for 24 blocks of trials, giving a total of 192 trials.

Test phase. The events on each trial of the test phase were the same as on a trial of the initial threshold estimation phase, with the exception that changes were not made to the patterns based on a participant's response. Each trial began with the presentation of a fixation cross for 1000ms. A pattern was then displayed on the left-hand side of the screen for 150ms followed by a masking display lasting for 250ms. Finally, a second pattern was displayed on the right-hand of the screen for 150ms and the computer then waited for the participant to respond by pressing either the 's' or 'd' key to indicate whether they believed that the two patterns were the same or different, respectively.

There was a total of 56 different types of test trial. These trial types were constructed by taking each pattern and comparing it with one of seven different patterns: itself, the three patterns that differed from it on a single dimension, and the three patterns that differed from it on two dimensions. Individual trial types were repeated different numbers of times within a block of trials to ensure that there was an equal number of trials on which the two patterns were the same, differed on a single dimension, or different on two dimensions. Furthermore, when the patterns differed on a single dimension, there was the same number of trials on which they differed on a relevant dimension as when they differed on the irrelevant dimension. Similarly, when the pattern differed on two dimensions, there was the same number of trials on which they differed on both relevant dimensions as when they differed on one relevant dimension and the irrelevant dimension. Hence, within a block of trials there were four trial on which a given pattern was compared with itself, one trial each on which it was compared with one of the patterns that differed from it on a single relevant dimension, two trials on which it was compared with the pattern that differed from it on the irrelevant dimension, two trials on which it was compared with the pattern that differed from it on both

relevant dimensions, and one trial each on which it was compared with one of the two patterns that differed from it on the irrelevant dimension and one of the relevant dimensions. Examples of these different trial types are shown in Figure 1a. There were 96 trials within a block and these were presented in a random order. The test phase consisted of two blocks of trials, for a total of 192 trials.

Results and Discussion

Because we were interested in the effects of category learning on perceptual similarity, we fully analysed the data only of participants who successfully solved the categorization task. A criterion of 75% correct responses over the final 64 trials of the categorization task was adopted for inclusion in data analysis. A total of 20 participants reached this criterion, and data from the remaining 20 participants were excluded from the statistical analyses reported below. On average, participant who reached the criterion made the correct response on 15.3 of the final 16 trials.

Threshold estimation. Participants required an average of 39.4 trials (range 25-66) to complete reversals on each staircase for the brightness dimension, 39.6 trials (range 21-88) for the horizontal dimension, and 36.8 trials (range 23-59) for the vertical dimension. A one-way analysis of variance (ANOVA) found no significant difference in the number of trials required to complete the staircases on the three dimensions, $F(2, 38) = 2.19$; $p = .126$. The group mean threshold values on the three dimensions were 25.2 units (range 4-40) of brightness, 9 pixels (range 4-22) of horizontal displacement, and 16 pixels (range 8-40) of vertical displacement. The latter two values corresponded to approximately 2.6 and 4.7 mm, respectively.

Category learning. Over the course of training, participants acquired the categorization task. The group mean proportion of correct responses is shown in Figure 4a for participants who reached the criterion of 75% correct responses over the last 64 trials, and for the participants who failed to reach this criterion. Figure 4b shows learning curves for those participants who reached the criterion, segregated according to which dimension was irrelevant to category membership. Participants for whom brightness was irrelevant learned the task most rapidly, and those for whom the vertical position of the line was irrelevant learnt the task least rapidly. By the end of training, however, participants in the three counterbalancing conditions had achieved comparable levels of performance. A two-way 3x12 mixed ANOVA with the between-subject factor of counterbalancing condition (brightness, horizontal, or vertical irrelevant) and the within-subject factor of trial block (1 to 12), confirmed these observations. There were significant effects of counterbalancing condition, $F(2, 17) = 4.55$, $p = .026$, mean square error (MSE) = .15, $\eta_p^2 = .35$, 90% CI[.28 .52], and of trial block, $F(11, 187) = 31.88$, $p < .001$, MSE = .01, $\eta_p^2 = .65$, 90% CI[.57 .68]. There was also a significant Condition x Block interaction, $F(22, 187) = 2.09$, $p < .005$, MSE = .01, $\eta_p^2 = .20$, 90% CI[.03 .19]. Simple effects analysis of the interaction revealed a significant effect of counterbalancing condition on trial blocks 2-6 and 8, smallest $F(2, 204) = 3.50$, $p = .032$. There was no significant effect of condition on any of the final four trial blocks across which the performance criterion was applied, final $F < 1$.

<Figure 4 about here>

Test phase. The group mean proportion of correct responses on trials in which the two patterns were the same (S), differed on the irrelevant dimension only (I), one relevant dimension (R), a relevant dimension and the irrelevant dimension (IR), and both relevant

dimensions (RR) across all patterns and both blocks of trials was 76.3% (SD = 11.3%), 42.3% (24.0%), 61.3% (13.1%), 72.8% (14.9%), and 80.5% (15.4%), respectively. Based on these values for individual participants, signal detection theory was then used to calculate each participant's sensitivity (*d'*) to each of the four types of difference. Average *d'* values are shown in Figure 5. Participants were more sensitive to the difference between two patterns when they differed on two dimensions than when they differed on one dimension. They were also more sensitive to the difference when the two patterns differed only on relevant dimensions than when there was a difference on an irrelevant dimension.

These observations were confirmed by a two-way 2x2 repeated measures ANOVA with the factors of number of different dimensions (one vs. two), and type of difference (relevant only vs. one irrelevant). The ANOVA revealed a significant effect of the number of different dimensions, $F(1, 19) = 66.95$; $p < .001$, MSE = .20, $\eta_p^2 = .78$, 90% CI[.58 .85] and of the type of difference, $F(1, 19) = 7.95$; $p = .011$, MSE = .52, $\eta_p^2 = .30$, 90% CI[.04 .50]. The interaction between these factors did not reach statistical significance, $F(1, 19) = 3.85$; $p = 0.064$.

<Figure 5 about here>

Using equations 2a and 2b, we were able to calculate that when two patterns differed on a dimension that was relevant in the categorization task, the *d'* measure of participants' sensitivity to the difference was 0.46 units higher than when the dimension was irrelevant. When two patterns belonged to different categories, *d'* scores were on average 0.13 units higher than when the two patterns belonged to the same category. One sample Student's *t*-tests found that the effect of attention was significantly different from zero, $t(19) = 2.82$; $p =$

.011, *d* = .63, 95% CI of mean [.11 .79]. The effect of mediation was not significantly different from zero, *t*(19) = 1.96; *p* = .064, *d* = .44, 95% CI [-.01 .26]. Finally, a pair-samples t-test confirmed that the effect of attention was greater than the effect mediation, *t(19)* = 2.49, *p* = .022, *d* = .56, 95% CI [.05 .60].

Replicated, metric, multi-dimensional scaling (MDS) analyses of the test data were performed to determine the extent to which the distances between stimuli had been modified in the manner illustrated in Figure 2a as a result of attentional weighting. *d'* scores were calculated using the proportion of correct responses made by each participant for each pair of stimuli presented during the test phase where the average accuracy across all 'same' trials was used to determine the false alarm rate; these scores provided a measure of the dissimilarity of two stimuli. Separate analyses were conducted for the three counterbalancing groups trained with a different irrelevant perceptual dimension. The ALSCAL command in SPSS (IBM Corp., Armonk, NY) was used to simultaneously analyse the dissimilarity matrices for all participants in each group, and solutions were requested in one, two, three, and four dimensions. The results from these analyses are shown in Figure 6. For all three groups, there was little improvement in stress when more than two dimensions were requested, and in each case the first two dimensions extracted corresponded to the perceptual dimensions that were relevant for that group of participants. The third dimension extracted corresponded to the perceptual dimension that was irrelevant for each group. The average distance between pairs of stimuli that differed on the irrelevant dimension was smaller than that between pairs of stimuli that differed on a relevant dimension. Importantly, four-dimensional solutions did not provide a better fit to the data than two- or three-dimensional solutions. In no case did

the analysis extract a dimension which corresponded to category membership which might be expected to result from a process of associative mediation.

<Figure 6 about here>

These results demonstrate that training on a conditional categorization tasks where category membership is determined by multiple stimulus dimensions resulted in a change in participants' sensitivity to differences between the stimuli. This extends the findings of Kruschke (1996) who found evidence for acquired distinctiveness effects following similar training, but using a two-stage transfer of training design. Furthermore, we found clear evidence for an effect of attentional weighting in acquired equivalence and distinctiveness, but at best marginal evidence for associative mediation.

Experiment 2

The purpose of Experiment 2 was to explore the generality of the results of Experiment 1 using a different category structure which also allowed a more direct measurement of the effect of attention than was possible in Experiment 1. The structure of the categorization task used in Experiment 2 is shown in Figure 1b. Rather than dividing the eight stimuli between two categories, they were distributed amongst four. As was the case in Experiment 1, two dimensions were relevant, and one dimension was irrelevant, to the categorization task. The key difference, however, was that stimuli which differed on two dimensions always belonged to different categories, regardless of whether both dimensions were relevant, or one of them was irrelevant.

As in Experiment 1, processes of attentional weighting and mediation were predicted to work in the same direction when two patterns differed on a single dimension. If that dimension was relevant, then the two patterns belonged to different categories. If the dimension was irrelevant, then they belonged to the same category. Hence, the difference in sensitivity to I and R differences should reflect the combined contribution of these two mechanisms. When the two patterns differed on two dimensions, however, in this experiment they always belonged to different categories. Hence, the contribution of associative mediation to a participant's sensitivity to the difference between such patterns should be equivalent regardless of whether both dimensions were relevant or only one was. Any difference in participants' sensitivity to IR and RR differences must, therefore, be due solely to the effects of attentional weighting. These relationships are reflected in Equations 3a and 3b:

$$R - I = attention + mediation \quad (3a)$$

$$RR - IR = attention \quad (3b)$$

By rearranging Equations 3a and 3b, we were able to calculate the separate contributions of attention and mediation to performance on the same-different task using Equations 4a and 4b:

$$attention = RR - IR \quad (4a)$$

$$mediation = (R - I) - (RR - IR) \quad (4b)$$

Method

Participants. A total of 40 undergraduate psychology students at the University of Hull served as participants. Before the start of the experiment they were randomly assigned in approximately equal numbers to three counterbalancing groups which differed in which stimulus dimension was irrelevant for the category learning task. Participants were compensated with either course credit (18) or a payment of £8 (22). Six of the participants failed to meet the learning criterion of 75% correct responses over the final 64 trials of the categorization task. All data for these participants were excluded from further analyses. Of the remaining 34 participants, 10, 10, and 14 participants belonged to the counterbalancing conditions in which brightness of the square, horizontal line position, and vertical line position, served as the irrelevant dimension in the categorization task, respectively. The ages of these 34 participants ranged from 18 to 29 years (*M* = 19.9), 23 were female, and 30 were right handed.

Stimuli and Apparatus. The stimuli and apparatus were the same as for Experiment 1.

Procedure.

Threshold estimation. This phase of the experiment proceeded in exactly the same manner as in Experiment 1.

Category learning. The structure of the categorization task used in this experiment is shown in Figure 1b. Eight patterns were created in the same manner as in Experiment 1 from the factorial combination of two values on each of the three perceptual dimensions on which the patterns varied. The category structure that was similar to that used in Experiment 1. Two

of the three dimensions were relevant to the solution of the categorization task whereas the third dimension was irrelevant. In this experiment, however, patterns were assigned to four, rather that two, categories. Hence, two patterns belonged to each category, and in each case patterns within these pairs differed only on the irrelevant dimension. The identity of the irrelevant dimension was different for each of the three counterbalancing groups.

Each trial began with the presentation of a white fixation cross in the centre of the screen. After 750ms, the fixation cross was replaced with one of the eight patterns, also in the centre of the screen. The caption 'Does this pattern belong to category F, G, H, or J?' was displayed above the pattern in 36pt white text. The caption remained on the screen until the participant responded by pressing one of the keys 'f', 'g', 'h', or 'j'. Following a correct response, the caption was replaced with the word 'Correct' in 60pt green text. Following an incorrect response, the caption was replaced with the word 'Incorrect' in 60pt red text. In either case, the correct answer was also presented; the words 'The pattern belongs to category [F/G/H/J].' were shown in 36pt white text below the pattern. This feedback remained on the screen for 2000ms after which the screen was cleared. There was then a 750ms interval before the next trial was initiated.

Within each block of eight trials, each of the eight patterns was presented once in a random order. Training continued for 24 blocks of trials, giving a total of 192 trials. The assignment of the labels 'f', 'g', 'h', and 'j' to the four categories to which the patterns belonged was determined randomly for each participant at the start of this phase of training.

Test phase. Details of the test phase are the same as for Experiment 1. There were a total of 56 different types of test trial constructed in same way as in Experiment 1. Example

of the comparisons made on these test trials are shown in Figure 1b. There were 96 trials within a block and these were presented in a random order. The test phase consisted of two blocks of trials, for a total of 192 trials.

Results and discussion

A criterion of 75% correct responses over the final 64 trials of the categorization task was again adopted for inclusion in data analysis. 34 participants reached this criterion, and data from the remaining 6 participants were excluded from all further statistical analysis. Over the final block of 16 trials, the participants that met the criterion were correct on average on 15.2 trials.

Threshold estimation. On average, participants required 40.9 trials (range 22-109) to complete reversals on each staircase for the brightness dimension, 37.2 trials (range 23-63) for the horizontal dimension, and 37.9 trials (range 21-77) for the vertical dimension. A one-way ANOVA found no reliable difference in the number of trial required to complete the staircases for the three dimensions, $F(2, 66) = 2.30$; $p = .109$. The group mean threshold values on the three dimensions were 27.1 units (range 11-40) of brightness, 7.4 pixels (range 4-14) of horizontal displacement, and 16.7 pixels (range 9-34) of vertical displacement. The latter two values corresponded to approximately 2.2 and 4.9 mm, respectively.

Category learning. Over the course of training, participants acquired the categorization task. The group mean proportion of correct responses is shown Figure 7a for participants who reached the criterion of 75% correct responses over the last 64 trials, and for the participants who failed to reach this criterion. Figure 7b shows learning curves for those participants who reached the criterion, segregated according to which dimension was

irrelevant to category membership. Participants for whom the horizontal position of the line was irrelevant appeared to perform slightly worse than participants in the other two counterbalancing conditions at the outset of training. This small difference disappeared rapidly, however, and the performance of the three groups was equivalent over the last two thirds of the training phase. A two-way 3x12 mixed ANOVA with the between-subject factor of counterbalancing condition (brightness, horizontal, or vertical irrelevant) and the within-subject factor of block of trials (1 to 12), confirmed these observations. There was a significant effect of trial block, $F(11, 341) = 51.77$, $p < .001$, MSE = .02, $\eta_p^2 = .63$, 90% CI[.56 .65]. There was no effect of counterbalancing condition, $F < 1$, and there was no significant Condition x Block interaction, $F(22, 341) = 1.52$, $p = .064$. A one-way ANOVA confirmed that there was no reliable difference between the counterbalancing groups on the final block of trials, $F < 1$.

<Figure 7 about here>

Test phase. The group mean proportion of correct responses on trials in which the two patterns were the same (S), differed on the irrelevant dimension only (I), one relevant dimension (R), a relevant dimension and the irrelevant dimension (IR), and both relevant dimensions (RR) across all patterns and both blocks of trials was 74.6% (14.8%), 41.4% (19.6%), 61.8% (21.8%), 66.5% (17.6%), and 79.6% (13.7%), respectively. Based on these values for individual participants, signal detection theory was then used to calculate each participant's sensitivity (*d'*) to each of the four types of difference in the same way as for Experiment 1. Average *d'* values are shown in Figure 8. Participants were more sensitive to the difference between two patterns when they differed on two dimensions than when they differed on one dimension. They were also more sensitive to the difference when the two

patterns differed only on relevant dimensions than when there was a difference on an irrelevant dimension.

<Figure 8 about here>

These observations were confirmed by a two-way 2x2 repeated measures ANOVA with the factors of number of different dimensions (one vs. two), and type of difference (relevant only vs. irrelevant). The ANOVA revealed a significant effect of the number of different dimensions, $F(1, 33) = 94.37$; $p < .001$, MSE = .16, $\eta_p^2 = .74$, 90% CI[.59 .81] and of the type of difference, $F(1, 33) = 20.22$; $p < .001$, MSE = .51, $\eta_p^2 = .38$, 90% CI[.16 .54]. The interaction between these factors did not reach statistical significance, $F(1, 33) = 1.83$; $P = 0.186$.

Using Equations 4a and 4b, estimates of the effects of attentional weighting and associative mediation were calculated. When two patterns differed on a dimension that was relevant in the categorization task, the *d'* measure of participants' sensitivity to the difference was 0.46 units higher than when the dimension was irrelevant. When two patterns belonged to different categories, *d'* scores were on average 0.18 units higher than when the two patterns belonged to the same category. One sample Student's *t*-tests found that the effect of attention was significantly different from zero, $t(33) = 4.09$; $p < .001$, $d = .70$, 95% CI of mean [.23 .70]. The effect of mediation was not different from zero, $t(33) = 1.35$; $p = .186$, $d = .23$, 95% CI [-.09 .44]. There was, however, no statistical difference between these two effects, $t(19) = 1.57$, $p = .127$, $d = .27$, 95% CI [-.09 .66].

The results of this experiment replicated those of Experiment 1, but using a slightly different category structure. Significant evidence was again found for an effect of attentional

weighting on the acquired equivalence and distinctiveness of stimuli, but there was no reliable evidence for an effect of associative mediation, although the difference between these effects was not significant.

Experiment 3

In Experiments 1 and 2, participants were trained on categorization tasks involving two relevant stimulus dimensions and one irrelevant dimension. In both experiments, participants learned to attend more to the relevant dimensions than to the irrelevant dimension, amplifying relevant differences relative to irrelevant differences. There was no evidence that associations between the stimuli and the category labels affected their discriminability to a significant degree. It is possible that associative mediation does affect discriminability in these categorization tasks, but that the effect is smaller than that of attentional weighting and, therefore, more difficult to detect. This might be especially true for the category structures used in these two experiments where the effect of associative mediation may only be observed as a second-order difference between conditions, and where it may have been overshadowed by much larger effects of attentional weighting.

The purpose of Experiment 3 was to try to isolate any potential effect of associative mediation by employing a categorization task in which it could be assessed as a first order difference between conditions, and where the effects of attentional weighting of stimulus dimensions might be expected to be smaller than in the previous experiments. We employed Shepard et al's (1961) type IV category structure in which all three stimulus dimensions were equally relevant to category membership, shown in Figure 1c. Each dimension was 75% diagnostic of category membership by itself, but the information from all three dimensions

was required to solve the categorization task fully. The structure of this tasks meant that comparisons could be made between patterns that differed on a single relevant stimulus dimension but that belonged either to the same, or to different categories (comparisons marked 1s and 1d, respectively, in Figure 1c). Similar comparisons were also possible for patterns that differed along two relevant dimensions (2s and 2d in Figure 1c). Because all three stimulus dimensions were relevant to the categorization task, effects of attention should be constant across comparisons between patterns differing on the same number of dimensions (i.e. 1s vs. 1d and 2s vs. 2d). Any variance in d' sensitivity to differences between pairs of patterns that belong to the same category and those that belong to different categories should, therefore, be attributable solely to processes of associative mediation. Hence, by subtracting participants' mean sensitivity to the difference between pairs of patterns that belong to the same categories (1s and 2s) from their mean sensitivity to the difference between pairs of patterns that belong to different categories (1d and 2d), we could obtain an estimate of the magnitude of the effect of mediation.

Method

Participants. Forty undergraduate psychology students at the University of Hull served as participants. Before the start of the experiment they were randomly assigned in approximately equal numbers to three counterbalancing groups. Participants were compensated with either course credit (20) or a payment of £8 (20). 23 of the participants failed to meet the learning criterion of 75% correct responses over the final 64 trials of the categorization task. All data for these participants were excluded from further analyses. The ages of the remaining 17 participants ranged from 18 to 33 years (*M* = 21.8), 12 were female, and 16 were right handed.

Stimuli and Apparatus. The stimuli and apparatus were the same as for Experiment 1.

Procedure.

Threshold estimation. This phase of the experiment proceeded in exactly the same manner as in Experiment 1.

Category learning. The structure of the categorization task used in this experiment is shown in the Figure 1c. Eight patterns were created in the same manner as in Experiment 1 from the factorial combination of two values on each of the three perceptual dimensions on which the patterns varied. The category structure was of a form consistent with that which Shepard et al (1961) referred to as a type IV classification. All three dimensions were relevant to the solution of the categorization task and each pattern was assigned to one of two categories. Three counterbalancing groups experienced slightly different versions of the type IV task. For the first group, the patterns ABX, APY, AQY, and BPY belonged to category 1 and the remaining patterns belonged to category 2 where A and B were the left and right horizontal positions of the bar, P and Q were the lower and upper vertical positions of the bar, and X and Y were the brighter and darker grey squares. Allocation of the labels ‘f’ and ‘g’ to categories 1 and 2 was determined randomly for each participant. For the other groups, the category structure was rotated. While these rotations did change the category membership of some patterns, it did not affect the relevance of each dimension. For the second group, patterns APX, BPX, BPY, and BQX were in category 1 and patterns APY, AQX, AQY, and BQY were in category 2. For the third group, patterns APX, AQX, AQY, and BQX were in category 1 and patterns APY, BPX, BPY, and BQY were in category 2.

Other details were the same as for Experiment 1. Within each block of eight trials, each of the eight patterns was presented once in a random order. Training continued for 24 blocks of trials, giving a total of 192 trials.

Test phase. Details of the test phase were similar to Experiments 1 and 2. Comparisons were made between pairs of patterns that differed on either one or two dimensions and belonged to either the same or different categories. Due to the structure of the categorization task, it was not possible to make all four of these comparisons for every pattern. Examination of Figure 1c reveals, for example, that for the first counterbalancing group any pattern to which APY differed on a single dimension (APX, AQY, and BPY) belonged to the same category as APY. Consequently, test trials were constructed so that all four types of comparison could be made against the first pattern presented on the trial. This meant that for each group, six of the eight patterns could be presented first on a test trial, with the same pattern and six of the other seven acting as comparison patterns (there were no test trials on which the two patterns differed on all dimensions). There was a total of 42 different types of test trials, examples of which are shown in Figure 1c. There were 72 trials within a block and these were presented in a random order. The test phase consisted of three blocks of trials, for a total of 216 trials.

Results and discussion

17 participants reached the criterion of 75% correct responses over the final 64 trials of the categorization phase. Data from the remaining 23 participants were excluded from all further statistical analysis. On average, participants who reached the criterion made the correct response on 14.8 of the final 16 trials. Examination of each participant's performance over the last four blocks of 16 trials suggested that none of them was relying on a simple uni-

dimensional rule; no such rule predicted performance across these four blocks. Instead, participants who reached the learning criterion did appear to be using information from all three stimulus dimensions.

Threshold estimation. On average, participants required 39 trials (range 27-62) to complete reversals on each staircase for the brightness dimension, 38.5 trials (range 20-70) for the horizontal dimension, and 37.9 trials (range 25-66) for the vertical dimension. A one-way ANOVA found no reliable difference in the number of trial required to complete the staircases for the three dimensions, $F < 1$. The group mean threshold values on the three dimensions were 26.9 units (range 12-38) of brightness, 11.6 pixels (range 4-35) of horizontal displacement, and 16.8 pixels (range 8-28) of vertical displacement. The latter two values corresponded to approximately 3.4 and 4.9 mm, respectively.

Category learning. Over the course of training, participants acquired the categorization task. The group mean proportion of correct responses is shown in Figure 9a for participants who reached the criterion of 75% correct responses over the last 64 trials, and for the participants who failed to reach this criterion. Because all three stimulus dimensions were relevant, and equally predictive of outcome for all participants, acquisition data are not shown separately for each counterbalancing group. There were, however, no reliable differences in the rates at which the three groups acquired the categorization task. This observation was confirmed by a two-way mixed ANOVA with the within-subject factor of trial block (1 to 12) and the between subject factor of counterbalancing group (1 to 3). This ANOVA reveal a significant effect of trial block, $F(11, 154) = 27.185$, $p < .001$, MSE = .01, $\eta_p^2 = .66$, 90% CI[.56 .69], but no effect of counterbalancing group, $F(2, 14) = 1.71$, $p = .217$, and no interaction between the factors, $F < 1$.

<Figure 9 about here>

Test phase. The group mean proportion of correct responses on trials in which the two patterns were the same (S), differed on a single dimension and belonged to the same category (1s), differed on a single dimension and belonged to different categories (1d), differed on two dimensions and belonged to the same category (2s), or differed on two dimensions and belonged to different categories (2d) across all patterns and both blocks of trials was 76.9% (13.7%), 58.6% (13.8%), 55.4% (14.1%), 78.2% (12.6%), and 76.6% (12.8%), respectively. Based on these values for individual participants, signal detection theory was then used to calculate each participant's sensitivity (*d'*) to each of the four types of difference in the same way as for Experiment 1. Average *d'* values are shown in Figure 9b. Participants were more sensitive to the difference between two patterns when they differed on two dimensions than when they differed on one dimension. Whether the two patterns belonged to the same category or to different categories had no effect on participants' sensitivity to the difference between them.

These observations were confirmed by a two-way 2x2 repeated measures ANOVA with the factors of number of different dimensions (one vs. two), and category membership (same vs. different). The ANOVA revealed a significant effect of the number of different dimensions, $F(1, 16) = 82.65$; $p < .001$, MSE = .01, $\eta_p^2 = .84$, 90% CI[.66 .89] but not of category membership, $F(1, 16) = 1.67$; $p = .215$, and no interaction between these factors, $F < 1$.

When two patterns belonged to different categories, *d'* scores were on average 0.08 units *lower* than when the two patterns belonged to the same category. A one sample

Student's *t*-tests found that this effect of associative mediation was not reliably different from zero, $t(16) = 1.29$; $p = .215$, $d = .31$, 95% CI of mean [-.22 .05].

For a third time, we failed to find any evidence for an effect of associative mediation on the discriminability of pairs of stimuli following training on a conditional categorization task. Under the null-hypothesis significance testing statistics so far reported, these null results are inconclusive. Bayesian methods, however, treat the experimental and null hypotheses as different possibilities, the evidence for each of which may be assessed. Hence, we may use Bayesian analysis to determine the likelihood of the null hypothesis – that associative mediation has no effect on the discriminability of stimuli in these experiments – given our results.

Bayesian Comparison of the Null and Experimental Hypotheses

Bayes factors may be calculated to give a measure of the relative strength of support for both experimental and null hypotheses. Values close to 1 provide no clear support in either direction, those above 3 provide substantial support for the experimental hypothesis, and those below 1/3 provide substantial support for the null hypothesis (Jeffreys, 1961; Dienes, 2011). We used a calculator provided by Dienes (2008) to compute Bayes factors for each experiment in turn. This calculator estimates the likelihood of the null and experimental hypotheses given the data. The likelihood of the null hypothesis is the height of the normal distribution with a mean of zero and standard deviation equal to the standard error of the sample mean. Since both attentional weighting and associative mediation are candidate mechanisms for an increase in sensitivity due to perceptual learning, directional predictions could be made in each case. This allowed us to model the prediction of each mechanism as a

half-normal distribution with a mean of zero and a standard deviation equal to the expected size of effect (Dienes, 2014). We estimated that the plausible size of effects for each mechanism of perceptual learning in Experiment 1 was unlikely to exceed 1 d' unit based on Goldstone's (1994) results in a related procedure (estimates of effects half this size, reflecting equal contribution of both mechanisms to Goldstone's acquired distinctiveness effect did not produce any substantial change in the results of the analysis). For analysis of Experiments 2 and 3, we were able to use the observed sizes of effects from Experiment 1 and 2, respectively, as estimates of the expected sizes of effects. Table 1 shows sizes of effects and Bayes ratios for the attention and mediation effects in each of the experiments.

<Table 1 about here>

Experiment 1 provided strong evidence, and Experiment 2 decisive evidence, in support of an effect of attentional weighting according to the categories proposed by Jeffreys (1961). The evidence that these two experiments each provided for an effect of associative mediation is, in Jeffreys' (p. 432) words, "not worth more than a bare mention". Experiment 3 provided substantial evidence against any effect of mediation in perceptual learning.

General Discussion

In three experiments, participants were trained on conditional categorization tasks involving stimuli which differed on three binary dimensions. Following this, their ability to discriminate between pairs of the stimuli was assessed using a same-different judgement task. In Experiments 1, two stimulus dimensions were relevant to category membership whereas the third dimension was irrelevant, and stimuli were assigned to one of two categories. This meant that when the stimuli differed on a single dimension, two candidate

mechanism of acquired equivalence and distinctiveness (associative mediation and attentional weighting), were expected to operate in the same direction. Each should have made a difference on a relevant dimension more easily detectable than a difference on an irrelevant dimension. When two stimuli differed on two dimensions, however, the two mechanisms were expected to operate in opposite directions. When both dimensions were relevant to the categorization task, changes in attentional weighting should have made the difference more noticeable than when one of the dimensions was irrelevant. A process of associative mediation would have had the opposite effect since stimuli differing on a single relevant dimension belonged to different categories and those differing on two relevant dimensions belonged to the same category. This allowed independent estimates to be made of the contributions of associative mediation and attentional weighting to acquired distinctiveness effects. We found a significant effect of attentional weighting, but only marginal evidence of associative mediation.

In Experiment 2, a slightly different category structure was employed where two stimulus dimensions were again relevant, and the third irrelevant, to the categorization task, but stimuli were divided into four, rather than two, categories. Here, a more direct measurement of the effect of attentional weighting was possible when two stimuli differed on two dimensions because any such pair of stimuli belonged to different dimensions. We again found a significant effect of attentional weighting, which was remarkably similar in magnitude to that observed in Experiment 1, but no reliable evidence of associative mediation. In Experiment 3, participants were trained on a categorization task where all three stimuli dimensions were relevant, allowing for a more direct measurement of any effect of associative mediation. Once more, we found no evidence of associative mediation.

Calculation of Bayes factors gave substantial support for the null hypothesis – that associative mediation does not contribute to acquired distinctiveness effects in these categorization tasks. The absence of a reliable effect of associative mediation is striking considering that our calculations assumed that the distance between stimuli in psychological space is determined by a city-block metric, and that the relationship between distance and (dis-)similarity is linear. These assumptions likely caused us to underestimate the magnitude of the effect of attentional weighting and overestimate that of associative mediation relative to either a Euclidean distance function, or an exponential decay function (e.g., Nosofsky, 1986) to relate distance and similarity. Hence, we might consider our estimates of the size of the effect of attentional weighing to be at its lower limit, and those of associative mediation to be at its upper limit.

Our calculations of the effects of attention and mediation in Experiments 1 and 2 were also based on the assumption that the influences of these two processes would summate. This assumption may be flawed; their interaction could be non-linear, or one process might dominate the other in particular circumstances. We shall return to this second possibility later, but for now we can at least assert that we found no evidence for an effect of associative mediation in the types of conditional categorization task that we tested. In Experiment 1, the difference in participants' sensitivity to I and R comparisons was the same as their sensitivity to IR and RR comparisons; the direction of the predicted effect of associative mediation had no reliable influence on test performance, summative or otherwise. In Experiment 2 we were able to directly assess the influence of attentional weighting. Again, associative mediation did not interact with this effect. Finally, in Experiment 3 we directly assessed associative

mediation in the absence of differences in attention weighting. For a third time, we failed to find evidence for associative mediation.

Despite our failure to find any evidence of associative mediation, it has been found to contribute to acquired distinctiveness effects in some situations. Hall et al (2003), reported results consistent with associative mediation in a series of experiments which controlled for the influence of attention. Meeter, Shohamy, and Myers (2009) also found evidence that associative mediation, rather than attention, contributes to acquired equivalence. Their participants learned to pair drawings of faces with different fishes, using a similar training problem as Hall et al; two faces were paired with one fish, and two other faces were paired with a second fish. For different groups of participants, different features of the faces (hair colour, gender, age) differed across, but not between the fish categories. On test trials, the hair colour of one of the faces was changed, and participants were required to select the fish that they though went with this new face. Because the pattern of responding was the same in all conditions, and favoured the identity of the face rather than hair colour, Meeter et al concluded that equivalence was governed by associative mediation via identity. They argued that changes in attentional weighting of the three features of the faces should result in more hair colour based responding when that feature was relevant than when it was irrelevant.

There are, however, a number of differences between these experiments and those reported here. The discrimination tasks were much simpler than our categorization tasks and, perhaps more importantly, the stimuli were much more discriminable than ours. It may be that under these conditions attentional weighting is less likely to be affected by learning than when differences between stimuli are more difficult to detect. Pothos and Reppa (2014) observed increases in within-category similarity following category learning on a difficult, but

not on an easy, problem where the stimuli were arrows which differed in height and width and task difficulty was manipulated by varying the separation of the stimuli along these dimensions. de Leeuw et al (2016) reported a similar effect of task difficultly on acquired equivalence and distinctiveness effects using more complex stimuli. Conversely, Pérez-Gay et al (2017) did not find an effect of stimulus discriminability on categorical perception effects, but their stimuli were textures generated by tessellation of simple checkerboard patterns, and it is arguable that even their most discriminable stimuli were substantially more similar than any of the stimuli used by Hall et al (2003) or by Meeter et al (2009). One explanation for this effect of discriminability might be found in the suggestion made by Sutherland and Mackintosh (1971) that attention to very salient stimuli might change more slowly than to less salient stimuli, a principle that has been incorporated into other models of associative learning (e.g., Suret & McLaren, 2003, 2005). Results consistent with this proposal were observed by Pearce, Esber, George, and Haselgrove (2008). They trained pigeons on a discrimination task where the stimuli consisted of geometric patterns and patches of colour. Faster learning was observed when colour was relevant than when the patterns were, suggesting that for pigeons colour is more salient than pattern. In a subsequent phase of the experiment, the pigeons' ability to discriminate between two patterns was worse for those for whom pattern had been irrelevant in the first phase of the experiment than for those for whom pattern had been relevant. The ability of the pigeons to discriminate between two colours was not affected by the relevance of colour in the original discrimination task. George (1998) found similar differences in the susceptibility of colour and spatial position to changes in attention.

It is also possible that while we found no evidence that similarity was mediated by simple associations between stimuli and category labels, more complex associative mechanism might explain out results. Honey and Ward-Robinson (2002) described a three-layer connectionist model of learning which, they claim (see also Honey, Close & Lin, 2010), can explain a range of acquired equivalence and acquired relational equivalence effects. In this network, units representing individual stimuli are connected to a layer of hidden units which are, in turn, connected bi-directionally to units representing the category labels (or outcomes). The strength of these connections is updated following each trial using Hebbian and anti-Hebbian learning algorithms. Robinson, George and Henke (2019) have recently presented a computational implementation of Honey and Ward-Robinson's network which confirmed many of Honey and colleague's earlier predictions about the behaviour of the model. We used the MATLAB code provided by Robinson et al to simulate the results of each of our experiments. In the absence of any attentional mechanism, the network model correctly predicted a greater effect of attentional weighting than of associative mediation for Experiment 1, but made the reverse prediction for Experiment 2 and incorrectly predicted a very substantial effect of associative mediation for Experiment 3 (see Appendix for details of the simulations).

Alternatively, effects attributed to associative mediation in human learning may, in fact, be the result of non-associative processes. Hall et al (2003) acknowledged that people may engage in verbal reasoning during their task. Having learned during phase 1 that "A goes with X" and during phase 2 that "A goes with left" they might reason that "X goes with left". In their fourth experiment, Hall et al found that participants actually believed that cues paired with the same antecedent prevented each other from appearing even though a group trained

to associate them with the same response learned faster than a group trained to associate them with different responses. It appeared that associative mediation determined performance despite participants' reasoning about the relationships between stimuli. Smyth, Barnes-Holmes and Barnes-Holmes (2008) argued, however, that participants may have re-evaluated the relationship between stimuli as preventative only when asked about them at test. In a series of experiments, they found that task instructions affected participants' evaluations of the relationships between stimuli, and that propositional knowledge was sufficient to produce acquired equivalence effects. These results are more consistent with a propositional model of associative learning (e.g., Lovibond, 2003), or Relational Frame Theory (Hayes, Barnes-Holmes & Roche, 2001), both of which explain acquired equivalence, and human contingency learning more generally, in terms of verbal processes. Liljeholm & Balleine (2010) also reported an acquired equivalence effect which cannot be explained in terms of either attentional weighting or associative mediation. The design of their experiment was again similar to that used by Hall et al, but during the first phase of training participants experienced numerous rapid contingency reversals. In a second phase of the experiment, learning was more rapid when participants had to learn the same response to stimuli that had received the same treatment, than if they had to learn different responses. Because each shape was equally often paired with each colour during phase one, this effect could not be due to shapes retrieving the memory of a colour which then became associated with a response (i.e. associative mediation). The stimuli were designed so that all features were either common to all shapes, or unique to one. This made it very unlikely that changes in attentional weighting of stimulus properties could account for the results, either. Liljeholm and Balleine conclude that acquired equivalence in their experiment must rely upon a

representation of the equivalence of relations between stimuli. Hence, their results are consistent with propositional models of learning. In our experiments, verbal reasoning might have been expected to produce the same effects as associative mediation; participants might have reasoned that stimuli which belonged to the same category 'go together' and as a result were more similar to each other. The absence of this effect might have been due to the complexity of the category structures and the similarity of the stimuli to each other; it was difficult to apply simple, distinctive labels to each of the eight stimuli. It might also have been due to the speeded testing procedure which left little time identify specific stimuli and then reason about their relationship to each other.

Goldstone, Lippa, and Shiffrin (2001) recognised that strategic judgement biases arising from verbal reasoning might affect similarity ratings following category learning. Having learned that two stimuli belong to the same category, participants might reason that they should be given a higher similarity rating than two stimuli which belong to different categories. Goldstone et al, however, devised a method of dissociating such effects of reasoning and those of attentional weighting. They trained participants on a categorization task similar to that used by Goldstone (1994), but where the stimuli were faces generated by morphing between different base photographs in such a way that they differed along two artificial dimensions. Before, and after, category learning, participants were given two types of test trials. In the first, they were asked to rate the similarity of two of the faces used in training which differed on either the irrelevant or the relevant stimulus dimension and, consequently, belonged to either the same category, or to different categories. As expected, the similarity of faces differing on the relevant dimension decreased as a result of category learning. This effect might be due to changes in stimulus representations (which could be

caused by either associative mediation or attentional weighting), or by strategic judgement biases. In the second type of test trial, participants were asked to rate the similarity of individual training stimuli to a neutral face. The difference in rating of similarity to the neutral face for stimuli belonging to the same category decreased following category learning. Goldstone et al suggested that this effect could not result from strategic judgement biases since the participants had not learned to attach any label to the neutral face. Instead, it could only be caused by changes to stimulus representations. As the representations of two stimuli become more similar to each other, their similarities to a third stimulus will become more consistent. That the two types of test trials revealed effects of training on different stimulus comparisons (within- vs. between-category) suggests that strategic judgement biases did contribute to similarity ratings between pairs of training stimuli. In our experiments, strategic judgement biases would be expected to operate in the same direction as any influence of associative mediation. The possibility that our participants might have developed such biases does not, therefore, affect our measurement of the effects of attentional weighting. If anything, our experiments will have overestimated the influence of associative mediation.

Although we found evidence only for attentional weighting, the existing evidence in favour of a process of associative mediation (if we accept that it is not explained by verbal reasoning) might be accommodated by dual-process models of category learning. COVIS (Ashby, Alfonso-Reese, Turken & Waldron, 1998; Ashby & Valentin, 2017) postulates two separate and competing learning systems and there are well-formed neurophysiological models of each process. A procedural system is mediated by the basal ganglia and can combine information from two or more dimensions, computing linear combinations of their dimensional values, or treating the stimulus as a gestalt. A declarative, rule-based system is

served by frontal brain regions and relies on working memory and executive attention. It can learn quickly when categories are separated by simple explicit rules. The two systems learn simultaneously and compete for control but learning of category structures which can be described by verbalizable rules will be dominated by the fast-learning rule-based system. Tasks which tap into the rule-based system may involve simple rules on a single stimulus dimension, but there is no reason why more complex rules cannot also be learned. Edmunds and Wills (2016) showed that the inclusion of conjunction and disjunction of conjunction rules allowed COVIS to solve the types of categorisation tasks that we employed and correctly predicted the ordering of all six category structures described by Shepard, Hovland and Jenkins (1961). Our results, then, might be explained by the learning of verbalizable rules (e.g., "line to the bottom left or top right is in category G, otherwise category H"), and switching of attention towards the rule-relevant perceptual dimensions. If associative mediation is an effect of the procedural learning system, it might not be expected to be present when learning is dominated by the declarative system. Conversely, we might assume that tasks of the type described by Hall et al (2003) favour the procedural learning system either because stimuli are arbitrarily paired with each other in phase one (shapes A and B → colour X; C and D → Y), or because phase two involves acquisition of motor responses. These assumptions may, however, be misplaced. First, because the effects attributed to associative mediation in human learning may result from verbal reasoning. Second, because Mathy et al (2013) found that performance on a type II categorization problem may be best described by a hybrid of exemplar- and rule-based models or by the network model SUSTAIN (Love, Medin & Gureckis, 2004), and not by a purely rule-based system.

The implications of our findings are twofold. First, they extend the range of categorization tasks in which acquired equivalence and distinctiveness effects have been observed. Second, the structures of these categorization tasks have also allowed us to separately assess the contribution of associative mediation and attentional weighting to these effects. We found considerable evidence for attentional weighting, but not for associative mediation. Since other authors have reported results consistent with associative mediation (or verbal reasoning), one question that remains to be answered is whether the structure of the task or the discriminability of the stimuli is a more important determinant of its influence on behaviour.

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*, 442-481.

Ashby, F. G., & Valentin, V. V. (2017). Multiple systems of perceptual category learning. In: H. Cohen & C. Lefebvre (Eds.), *Handbook of Categorization in Cognitive Science* (2nd ed., pp. 547–572). New York, NY: Elsevier.

Bende, M., & Nordin, S. (1997). Perceptual learning in olfaction: professional wine tasters versus controls. *Physiology & Behavior, 62*, 1065-1070.

Biederman, I., & Shiffrar, M. M. (1987). Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual-learning task. *Journal of Experimental Psychology: Learning, Memory and Cognition, 13*, 640-645.

Bonardi, C., Graham, S., Hall, G., & Mitchell, C. (2005). Acquired equivalence and distinctiveness in human discrimination learning: Evidence for an attentional process. *Psychonomic Bulletin & Review, 12*, 88-92.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433-436.

de Leeuw, J. R., Andrews, J. K, Livingstone, K. R., & Chin, B. M. (2016). The effects of categorization on perceptual judgment are robust across different assessment tasks. *Collabra, 2(1): 9*, 1–9.

Dienes, Z. (2008). *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference.* Basingstoke: Palgrave Macmillan.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6*, 274-290.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781.

Edmunds, C. E. R., & Wills, A. J. (2016). Modeling category learning using a dual-system approach: A simulation of Shepard, Hovland and Jenkins (1961) by COVIS*. Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review, 57*, 94–104.

George, D. N. (1998). *Acquire distinctiveness*. (Unpublished doctoral dissertation). Cardiff University, Cardiff, Wales.

George, D. N., & Pearce, J. M. (1999). Acquired distinctiveness is controlled by stimulus relevance not correlation with reward. *Journal of Experimental Psychology: Animal Behavior Processes, 25*, 363-373.

Gibson, E. J. (1969). *Principles of perceptual learning and development*. New York: Appleton-Century-Crofts.

Gibson, J. J., & Gisbon, E. J., (1955). Perceptual learning: Differentiation or enrichment? *Psychological Review, 62*, 32-41.

Goldstone, R. (1994). Influences of categorization on perceptual learning. *Journal of Experimental Psychology: General, 123*, 178-200.

Goldstone, R. (1998). Perceptual learning. *Annual Review of Psychology, 49*, 585-612.

Goldstone, R. (2003). Learning to perceive while perceiving to learn. In R. Kimchi, M. Behrmann, & C. Olson (Eds.) *Perceptual organization in vision: Behavioral and neural perspectives.* (pp. 233-278). New Jersey: Lawrence Erlbaum Associates.

Goldston, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition, 78*, 27-43.

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General, 130*, 116-139.

Hall, G. (1991). *Perceptual and associative learning*. Oxford: Oxford University Press.

Hall, G., Mitchell, C., Graham, S., & Lavis, Y. (2003). Acquired equivalence and distinctiveness in human discrimination learning: Evidence for associative mediation. *Journal of Experimental Psychology: General, 132*, 266-276.

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). *Relational frame theory: A post-Skinnerian account of human language and cognition*. New York: Plenum Press.

Honey, R. C., Close, J., & Lin, T. E. (2010). Acquired distinctiveness and equivalence: A synthesis. In Mitchell, C. J. & Le Pelley, M. E. (Eds.), *Attention and associative learning: From brain to behaviour* (pp. 159–186). Oxford: Oxford University Press.

Honey, R. C., & Ward-Robinson, J. (2002). Acquired equivalence and distinctiveness of cues: I. Exploring a neural network approach. *Journal of Experimental Psychology: Animal Behavior Processes, 28*, 378–387.

Hull, C. L. (1939). The problem of stimulus equivalence in behavior theory. *Psychological Review, 46*, 9-30.

James, W. (1890). *The principles of psychology*, in two volumes. New York, NY: Henry Holt and Company.

Jeffreys, H. (1961). *The theory of probability* (3rd ed.). Oxford, England: Oxford University Press.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception*, **36**, ECVP Abstract Supplement.

Konorski, J. (1948). *Conditioned reflexes and neuron organization*. Cambridge: Cambridge University Press.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22-44.

Kruschke, J. K. (1996). Dimensional relevance shifts in category learning. *Connection Science, 8*, 201-223.

Kruschke, J. K. (2001a). The inverse base-rate effect is not explained by eliminative interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 1385-1400.

Kruschke, J. K. (2001b). Towards a unified model of attention in associative learning. *Journal of Mathematical Psychology, 45*, 812-863.

Lawrence, D. H. (1949). Acquired distinctiveness of cues: I. Transfer between discriminations on the basis of familiarity with the stimulus. *Journal of Experimental Psychology, 39*, 770-784.

Lawrence, D. H. (1950). Acquired distinctiveness of cues: II. Selective association in a constant stimulus situation. *Journal of Experimental Psychology, 40*, 175-188.

Lawrence, D. H. (1952). The transfer of a discrimination along a continuum. *Journal of Comparative and Physiological Psychology, 45*, 511-516.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology, 54*, 358-368.

Liljeholm, M., & Balleine, B. (2010). Extracting functional equivalence from reversing contingencies. *Journal of Experimental Psychology: Animal Behavior Processes, 36*, 165-171.

Livingstone, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 732-753.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111*, 309-332.

Lovejoy, E. (1968). *Attention in discrimination learning*. San Francisco: Holden-Day.

Lovibond, P. F. (2003). Causal beliefs and conditioned responses: Retrospective revaluation induced by experience and by instruction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 97–106.

Lupyan, G. (2012). Linguistically modulated perception and cognition: the label-feedback hypothesis. *Frontiers in Psychology, 3*, 54.

Lupyan, G., Rakson, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels Facilitate learning of novel categories. *Psychological Science, 18*, 1077-1083.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82*, 276-298.

Mathy, F., Haldjian, H. H., Laurent, E., & Goldstone, R. L. (2013). Similarity-dissimilarity competition in disjunctive classification tasks. *Frontiers in Psychology, 4*, 26.

Meeter, M., Shohamy, D., & Myers, C. E. (2009). Acquired equivalence changes stimulus representations. *Journal of the Experimental Analysis of Behavior, 91*, 127-141.

Miller, N. E. (1948). Theory and experiment relating psychoanalytic displacement to stimulus-response generalization. *Journal of Abnormal and Social Psychology, 43*, 155-178.

Miller, N. E., & Dollard, J. (1941). *Social learning and imitation*. New Haven, CT: Yale University Press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57.

Notman, L. A., Sowden, P. T., & Özgen, E. (2005). The nature of categorical perception effects: A psychophysical approach. *Cognition, 95*, B1-B14.

Pearce, J. M., Esber, G. R., George, D. N., & Haselgrove, M. (2008). The nature of discrimination learning in pigeons. *Learning & Behavior, 36,* 188-199.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437-442.

Pérez-Gay, F., Thériault, C., Gregory, M., Sabri, H., Rivas, D., & Harnad, S. (2017). How and why does category learning cause categorical perception? *International Journal of Comparative Psychology,* 30.

Peron, R. M., & Allen, G. L. (1988). Attempts to train novices for beer flavour discrimination: A matter of taste. *Journal of General Psychology, 115*, 403-418.

Pothos, E. M., & Reppa, I. (2014). The fickle nature of similarity change as a result of categorization. *Quarterly Journal of Experimental Psychology, 67*, 2425-2438.

Quinlan, P. T. (2010). On the use of the term 'attention'. In C. J. Mitchell and M. E. Le Pelley (Eds.), *Attention and Associative Learning: From Brain to Behaviour*. Oxford, UK: Oxford University Press.

Robinson, J., George, D. N., & Heinke, D. (2019). A computational implementation of a Hebbian learning network and its application to configural forms of acquired equivalence. *Journal of Experimental Psychology: Animal Learning and Cognition, 45*, 356-371.

Shapiro, P. N., & Penrod, S. (1986). Meta-analysis of facial identification studies. *Psychological Bulletin, 100*, 139-156.

Shepard, R. N., Hovland, H. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75* (13, Whole Number 517).

Smyth, S., Barnes-Holmes, B., & Barnes-Holmes, Y. (2008). Acquired equivalence in human discrimination learning: The role of propositional knowledge. *Journal of Experimental Psychology: Animal Behavior Processes, 34*, 167-177.

Sowden, P. T., Davies, I. R. L., & Roling, P. (2000). Perceptual learning of the detection of features in X-ray images: A functional role for improvements in adults' visual sensitivity? *Journal of Experimental Psychology: Human Perception and Performance, 26*, 379-390.

Suret, M., & McLaren, I. P. L. (2003). Representation and discrimination on an artificial dimension. *Quarterly Journal of Experimental Psychology, 56B,* 30-42.

Suret, M., & McLaren, I. P. L. (2005). Elemental representation and associability: An integrated model. In A.J. Wills (Ed.), *New Directions in Human Associative Learning* (pp. 155-187). New York: Psychology Press.

Sutherland, N. S., & Mackintosh, N. J. (1971). *Mechanisms of animal discrimination learning*. New York: Academic Press.

Walk, R. D. (1966). Perceptual learning and the discrimination of wines. *Psychonomic Science, 5*, 57-58.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences, 104*, 7780-7785.

Zeaman, D., & House, B. J. (1963). The role of attention in retardate learning. In N. R. Ellis (Ed.), *Handbook of mental deficiency: Psychological theory and research* (pp. 159-223). New York: McGraw-Hill.

Author Note

This research was supported by a Small Grant from the Experimental Psychology Society (eps.ac.uk) and is based on ideas developed when DNG was a visiting research fellow at the School of Psychology, UNSW Australia. DNG is grateful to UNSW Australia for the generous provision of facilities during this period.

Table 1

Bayesian *t*-test analyses of the effects of attention weighting and associative mediation in each experiment. Bayes factors for Experiment 1 are given for both large (and small) estimated sizes of effects. *strong support for the experimental hypothesis, **decisive support for the experimental hypothesis, †substantial support for the null hypothesis.

	Mean, d'	*Effect size, d*	*t-statistic*	*p-value*	*Bayes factor, B*
Experiment 1					
Attention	.46	.63	2.82	.011	15.3* (22.35*)
Mediation	.13	.44	1.96	.064	.86 (1.66)
Experiment 2					
Attention	.46	.70	4.09	<.001	1261**
Mediation	.18	.23	1.35	.186	1.85
Experiment 3					
Mediation	-.08	-.31	1.29	.215	.16†

Figure Captions

Figure 1. The structure of the categorization tasks used in (a) Experiment 1, (b) Experiment 2, and (c) Experiment 3. The three stimulus dimensions are represented as the three spatial dimensions of the cubes. Circles at each vertex of a cube represent the eight individual stimuli. Category membership is indicated by the colour of the circle. Comparisons made on test trials are indicated relative to stimulus APX (S), at the front, bottom, left corner of each cube.

Figure 2. Representation of the hypothetical distances between stimuli following category learning in Experiment 1 as a result of (a) changes in attentional weighting where the stimuli are stretched apart along the relevant perceptual dimensions and/or compressed on the irrelevant dimension, and (b) associative mediation which increases differences between members of different categories and/or reduces difference between members of the same category.

Figure 3. The prototypical pattern (a) from which all experimental stimuli were derived, and the sequence of events on each trial of the same-different judgement task (b). The dashed rectangles show the areas within which the stimuli could be presented, but these rectangles did not appear on the screen. Colours have been inverted; all displays used a black background and the vertical line in each pattern was white.

Figure 4. Proportion of correct responses on each of the 12 blocks of 16 trials of the categorization phase of Experiment 1. (a) Average performance of participants who reached the criterion of at least 75% correct responses over the final four blocks of trials (n = 20), and of those participants that did not (n = 20). (b) Average performance of the participants who

reached the criterion in each of the three counterbalancing conditions in which a different stimulus dimension was irrelevant. The reference line indicates chance performance (50% correct). Error bars show one standard error of the mean (1 SEM).

Figure 5. Average d' sensitivity to differences between pairs of patterns compared during the test phase of Experiment 1. Patterns could differ on one relevant (R), one irrelevant (I), two relevant (RR), or one relevant and one irrelevant (IR) dimension(s). Error bars show 1 SEM.

Figure 6. Two- (a, c, e), and three-dimensional (b, d, f) multi-dimensional scaling (MDS) solutions to dissimilarity measures derived from the test data in Experiment 1 for participants for whom brightness (a and b), horizontal line position (c and d), or vertical line position (e and f) was irrelevant in the categorization task. Axes are labelled with the perceptual dimension which most closely matches the extracted dimension. For each group of participants, a scree plot shows the stress for solutions in one to four dimensions. Numbers presented to the top-right of each plot give, for each perceptual dimension, the average distance between the four pairs of stimuli that differ only on that dimension (e.g., aPX vs. bPX; aPY vs. bPY; aQX vs. bQX; aQY vs. bQY). Upper-case letter indicate that the perceptual dimension was relevant to the categorization task, lower-case letters indicate that it was irrelevant. Red lines join the midpoints between the four vertices which share a value on one dimension (e.g., aPX, aPY, aQX, aQY vs. bPX, bPY, bQX, bQY) and provide an indication of both how well separated the perceptual dimensions are in each MDS solution, and also how well separated are the values along each dimension. Category membership of stimuli is indicated by blue and white circles at the vertices.

Figure 7. Proportion of correct responses on each of the 12 blocks of 16 trials of the categorization phase of Experiment 2. (a) Average performance of participants who reached the criterion of at least 75% correct responses over the final four blocks of trials (n = 34), and of those participants that did not (n = 6). (b) Average performance of the participants who reached the criterion in each of the three counterbalancing conditions in which a different stimulus dimension was irrelevant. The reference line indicates chance performance (25% correct). Error bars show 1 SEM.

Figure 8. Average d' sensitivity to differences between pairs of patterns compared during the test phase of Experiment 2. Patterns could differ on one relevant (R), one irrelevant (I), two relevant (RR), or one relevant and one irrelevant (IR) dimension(s). Error bars show 1 SEM.

Figure 9. (a) Proportion of correct responses on each of the 12 blocks of 16 trials of the categorization phase of Experiment 3 for participants who reached the criterion of at least 75% correct responses over the final four blocks of trials (n = 17), and for those participants that did not (n = 23). (b) Average d' sensitivity to differences between pairs of patterns compared during the test phase of Experiment 3. Patterns could differ on one or two relevant dimensions and belong to either the same category or to different categories. The reference line indicates chance performance (50% correct). Error bars show 1 SEM.

Figure 1

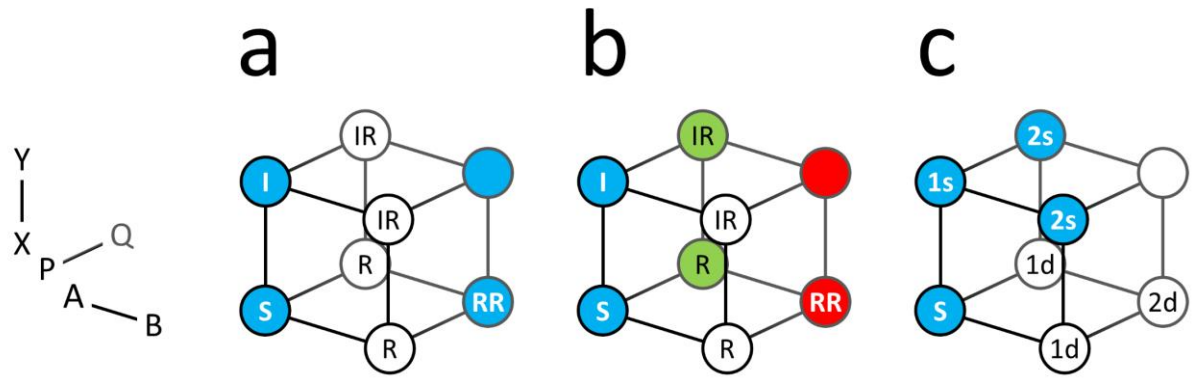


Figure 2

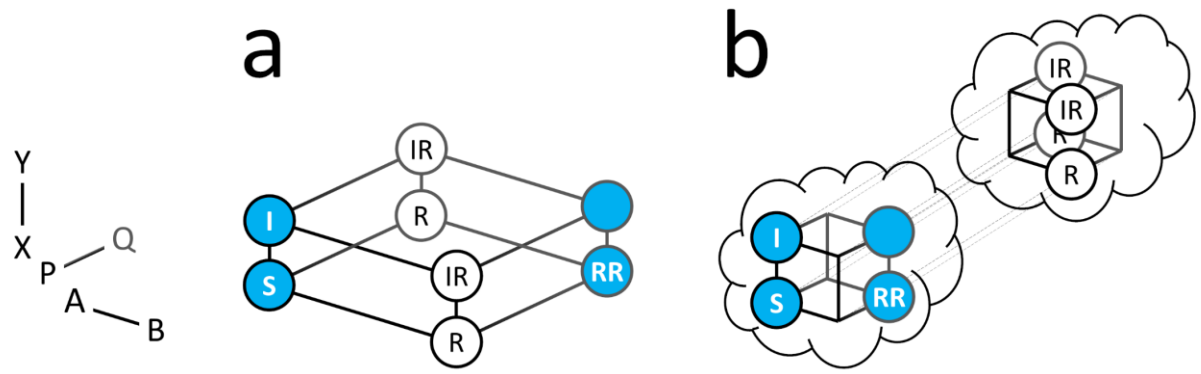


Figure 3

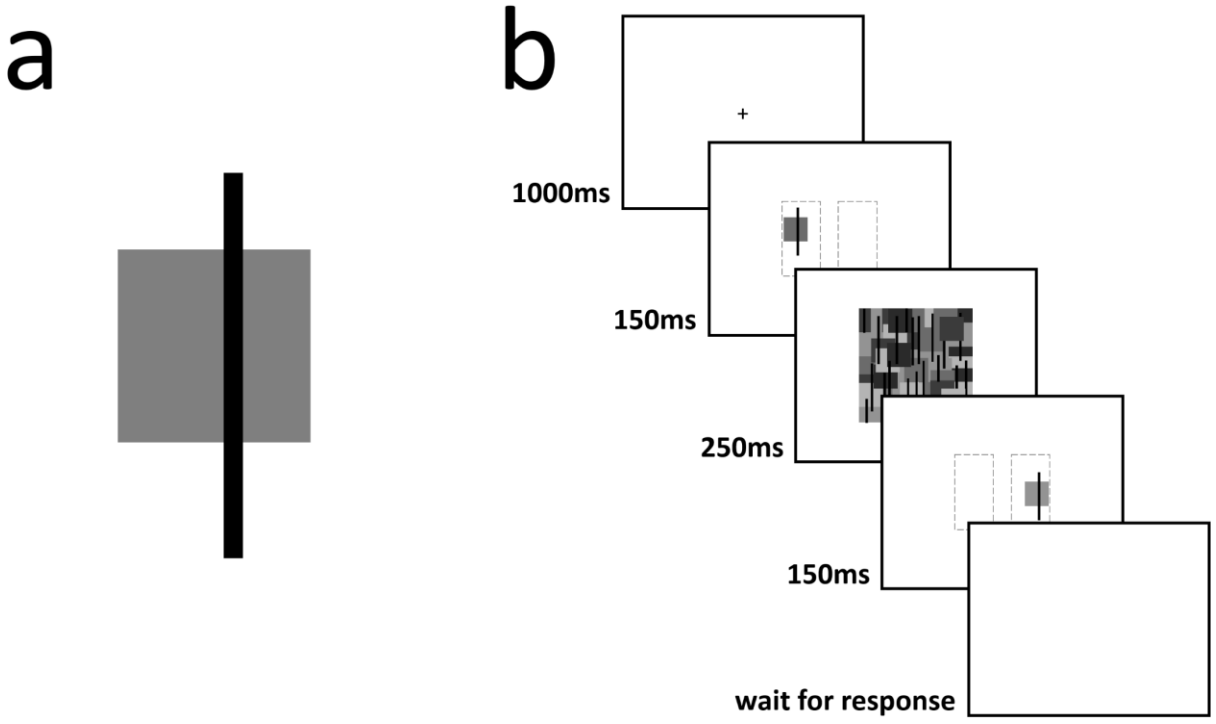


Figure 4

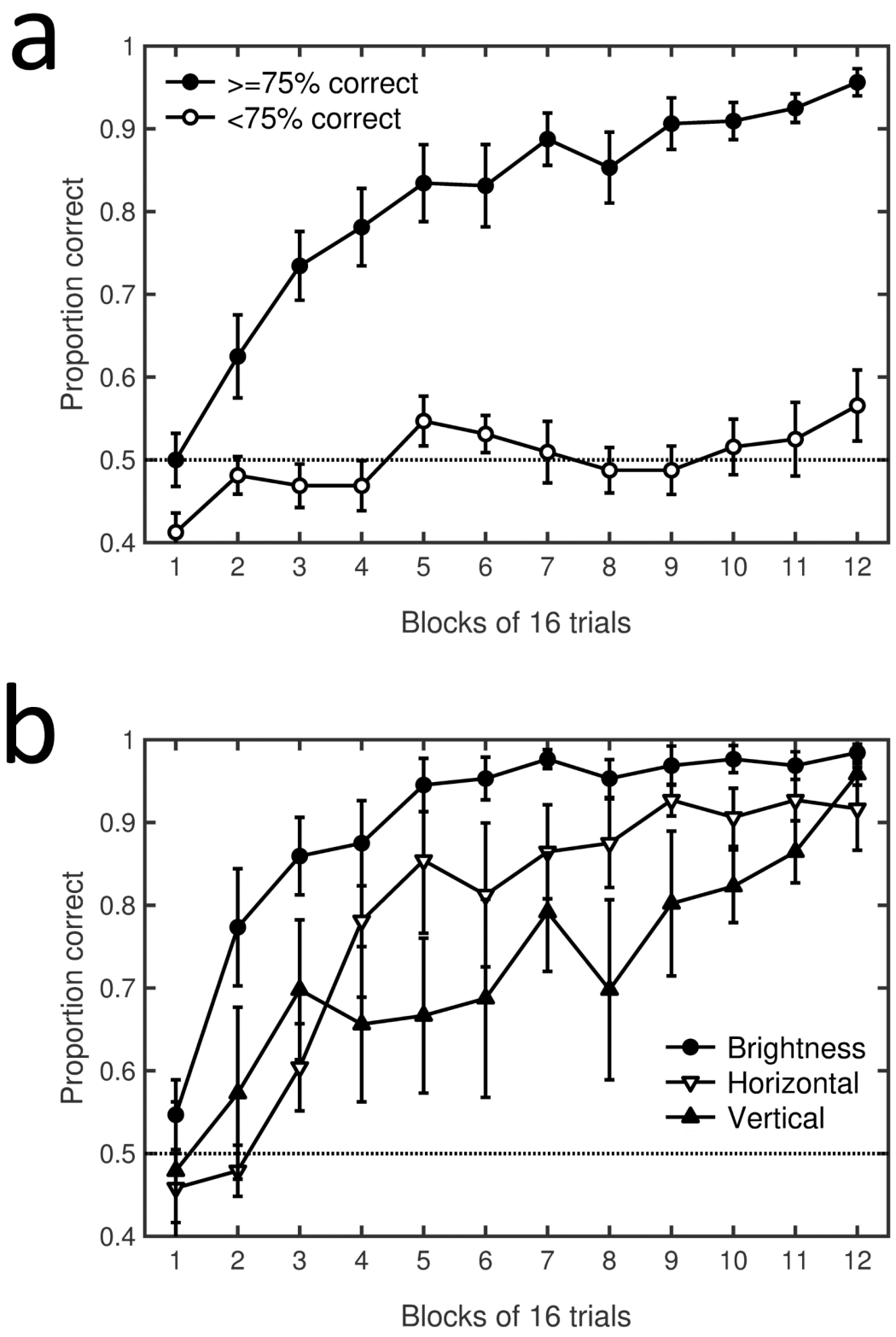

a
>=75% correct
<75% correct
Proportion correct
Blocks of 16 trials
b
Brightness
Horizontal
Vertical
Proportion correct
Blocks of 16 trials

Figure 5

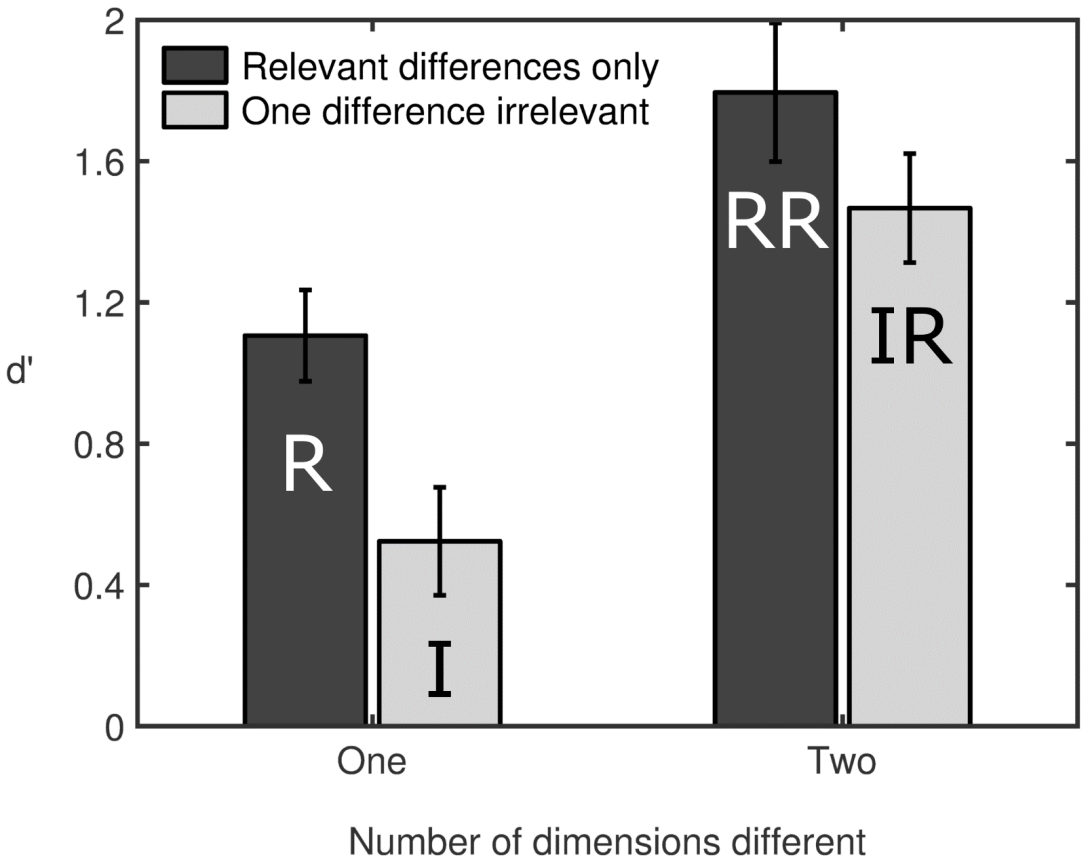

Relevant differences only
One difference irrelevant
R
I
RR
IR
d'
2
1.6
1.2
0.8
0.4
0
One
Two
Number of dimensions different

Figure 6

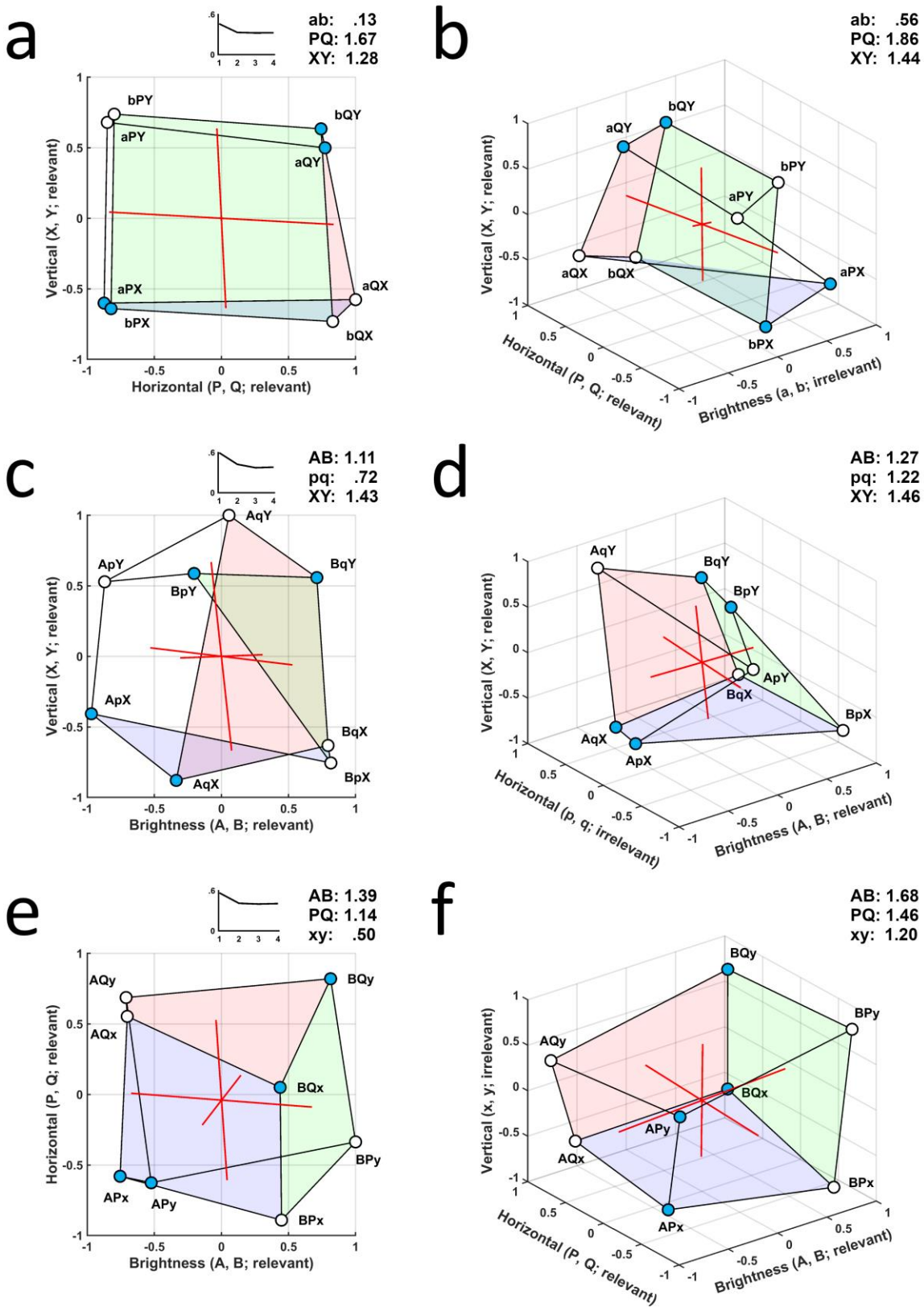

a
ab: .13
PQ: 1.67
XY: 1.28
bPY
bQY
aPY
aQY
aPX
aQX
bPX
bQX
Vertical (X, Y; relevant)
Horizontal (P, Q; relevant)
b
ab: .56
PQ: 1.86
XY: 1.44
bQY
aQY
bPY
aPY
aQX
bQX
aPX
bPX
Vertical (X, Y; relevant)
Horizontal (P, Q; relevant)
Brightness (a, b; irrelevant)
c
AB: 1.11
pq: .72
XY: 1.43
AqY
ApY
BqY
BpY
ApX
BqX
BpX
AqX
Vertical (X, Y; relevant)
Brightness (A, B; relevant)
d
AB: 1.27
pq: 1.22
XY: 1.46
AqY
BqY
BpY
ApY
BqX
BpX
AqX
ApX
Vertical (X, Y; relevant)
Horizontal (p, q; irrelevant)
Brightness (A, B; relevant)
e
AB: 1.39
PQ: 1.14
xy: .50
AQy
BQy
AQx
BQx
BPy
APx
APy
BPx
Horizontal (P, Q; relevant)
Brightness (A, B; relevant)
f
AB: 1.68
PQ: 1.46
xy: 1.20
BQy
BPy
AQy
BQx
APy
AQx
BPx
APx
Vertical (x, y; irrelevant)
Horizontal (P, Q; relevant)
Brightness (A, B; relevant)

Figure 7

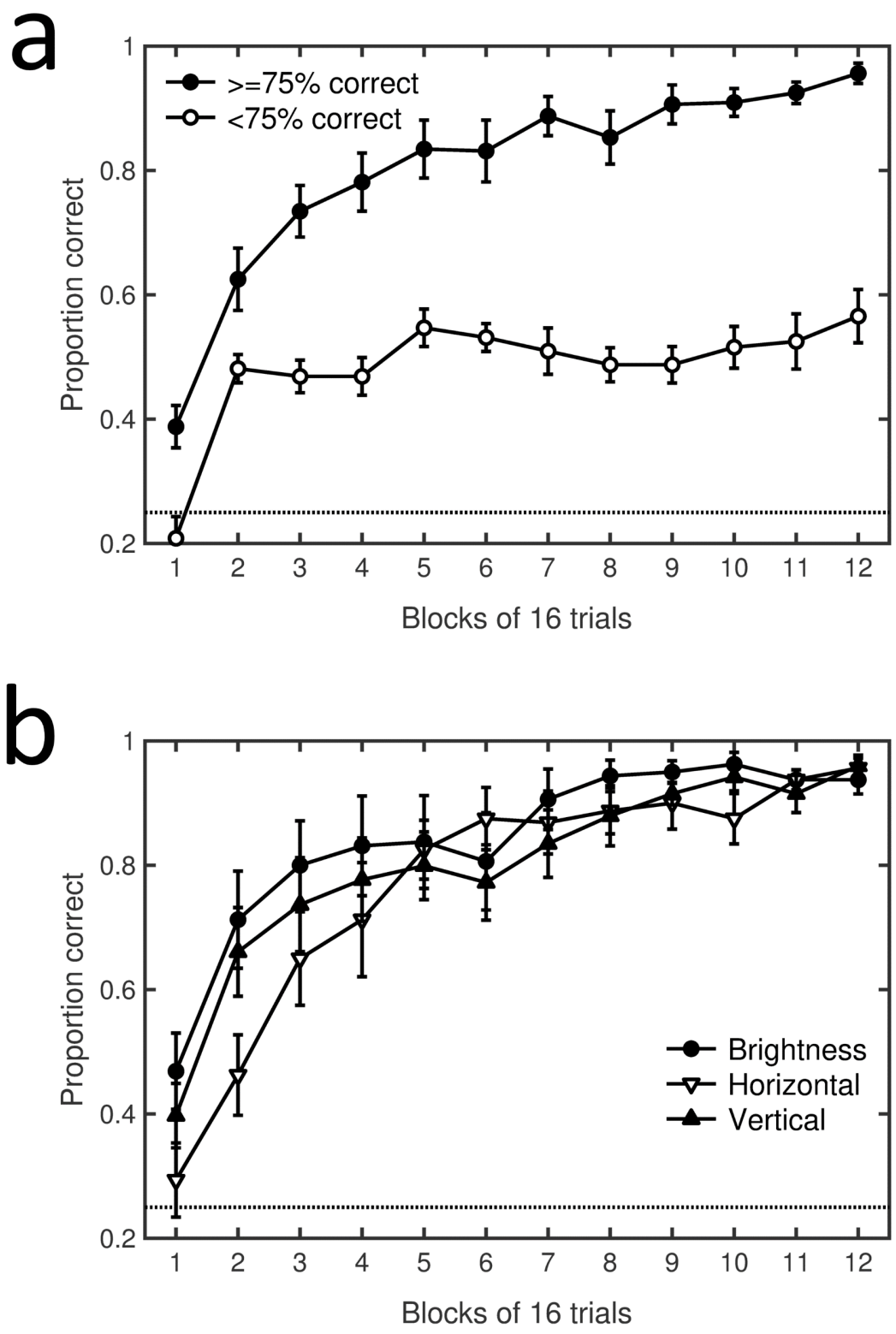

a
>=75% correct
<75% correct
Proportion correct
Blocks of 16 trials
b
Brightness
Horizontal
Vertical
Proportion correct
Blocks of 16 trials

Figure 8

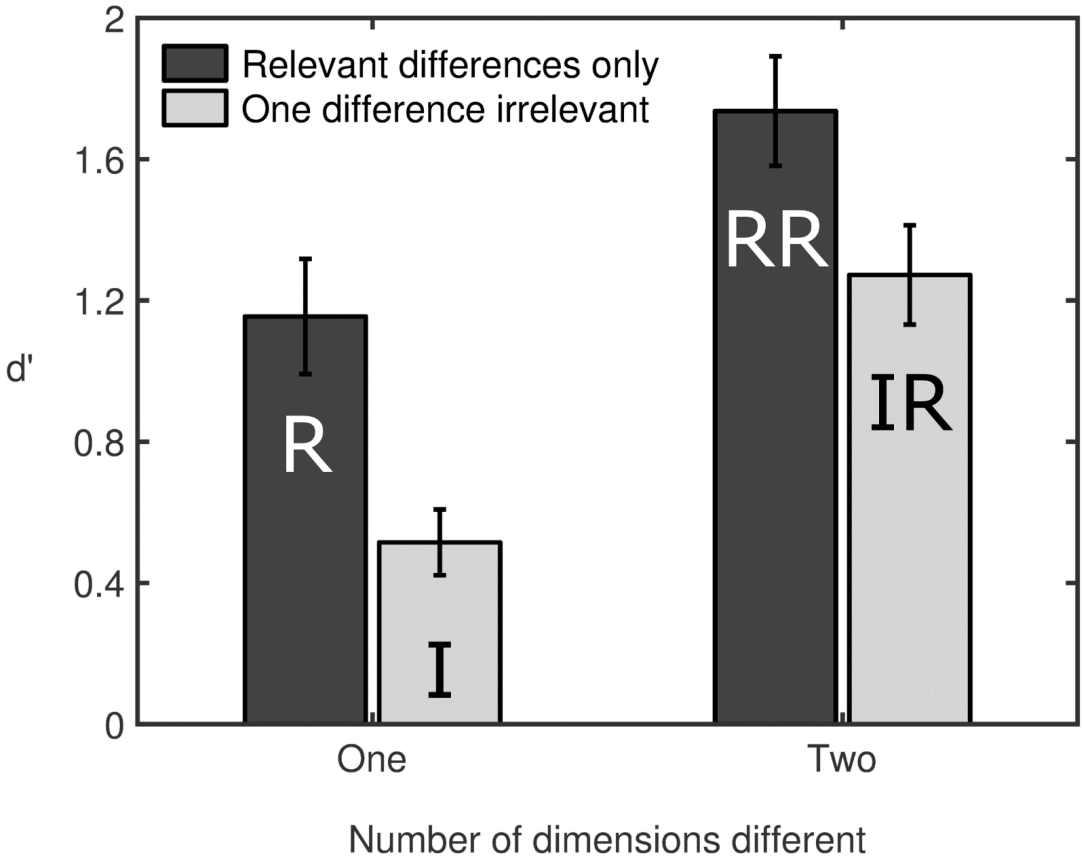
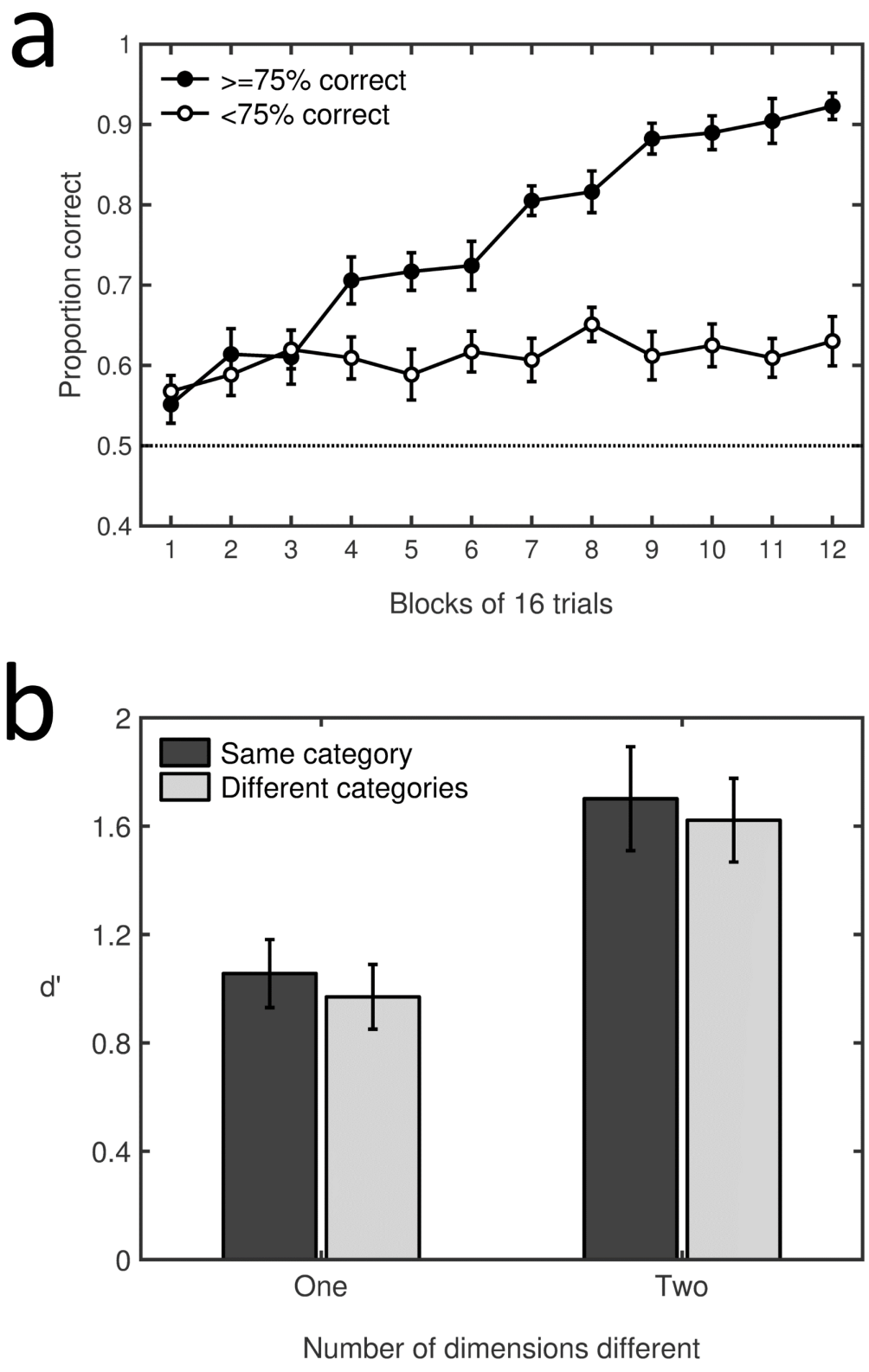

Relevant differences only
One difference irrelevant
d'
2
1.6
1.2
0.8
0.4
0
R
I
RR
IR
One
Two
Number of dimensions different

Figure 9


a
>=75% correct
<75% correct
Proportion correct
1
0.9
0.8
0.7
0.6
0.5
0.4
1 2 3 4 5 6 7 8 9 10 11 12
Blocks of 16 trials
b
Same category
Different categories
d'
2
1.6
1.2
0.8
0.4
0
One
Two
Number of dimensions different

Appendix

Details of simulations of Honey and Ward-Robinson's (2002) connectionist model of acquired equivalence

Simulations of Robinson et al's (2019) implementation of Honey and Ward-Robinson's (2002) network model were conducted using the MATLAB code available from the "HebbianNN" repository on GitHub (GitHub Inc., San Francisco, CA) at https://github.com/DavidNGeorge/HebbianNN and on the "A Computational Implementation of a Hebbian Learning Network and Its Application to Configural Forms of Acquired Equivalence" repository on American Psychological Association's Open Science Framework at https://osf.io/sjbd6/.

For each experiment, repeated simulations of the network were run with different random starting values for the weights between units. Each network consisted of six input units which represented the two stimulus values on each of three stimulus dimensions ([A, B]; [P, Q]; and [X, Y]), four hidden units, and either two or four output units corresponding to the categories. Additional simulations run with eight hidden units yielded very similar results. Each stimulus was represented at the input layer by a six binary-digit vector (e.g., APX = [1 0 1 0 1 0]; BQW = [0 1 0 1 0 1]). Similarly, categories were represented by setting the value of a single output unit to 1 and of the other output unit(s) to 0. Each network was trained for 50 epochs, in each of which the eight stimuli were each presented once in a random order. The learning rate parameter was set at .1 for all weights. All other parameters were set at the default values described by Robinson et al If the average root mean square error across the output units for all training stimuli fell below an arbitrary criterion of .2 by the end of training,

the network was considered to have solved the categorization problem and was then tested. For each experiment a total of 10,000 successful simulation runs were completed.

The pattern of activity across the hidden units in response to the presentation of a stimulus comprises the network's internal representation of that stimulus. Hence, at the end of training, the same/different judgement task was simulated by presenting two patterns to the network and comparing the resulting patterns of hidden unit activity. The greater the disparity between these patterns, the greater the ability of the network to discriminate between the stimuli. The discriminability of two stimuli was operationalized by calculating the quadratic mean of the difference in activation of each hidden unit in response to the two patterns (i.e., the root mean square difference). This number was in the range 0-1 where higher values corresponded to greater discriminability. Because random noise was added to the activation of each unit in the network on every trial, each comparison test trial was repeated 1000 times with different patterns of noise across the network. The mean discriminability of pair of patterns which corresponded to each of the types of comparisons shown in Figure 1 was then calculated. These values were used to compute the observed effects of attentional weighting and associative mediation.

Simulations of the categorization task used in Experiment 1 revealed that the predicted mean discriminability of two patterns that differed only on the irrelevant dimension (I) was .00; of patterns that differed only on one of the relevant dimensions (R) was .56; of patterns that differed on the irrelevant dimension and one of the irrelevant dimensions (IR) was .56; and of patterns than differed on both relevant dimensions (RR) was .69. From Equations 2a and 2b, the calculated effect of attentional weighting was .34 and of associative

mediation was .21. Hence, the predictions of the network provided a fairly good match to the results that we observed in Experiment 1.

For simulations of Experiment 2, the discriminability of the patterns of hidden unit activity for the four type of comparison were: I = .00, R = .61, IR = .61, RR = .71. From Equations 4a and 4b, the calculated effect of attentional weighting was .09 and of associative mediation was .52. The order of these predictions is opposite to the effects that we observed in Experiment 2.

Simulations of Experiment 3 found that the discriminability of patterns differing on a single dimension and belong to the same (1s), or different (1d) categories was .19 and .68, respectively. The discriminability of patterns differing on two dimensions and belong to the same (2s), or different (2d) categories was .05 and .83, respectively. These values resulted in a very large predicted effect of mediation of .63, whereas Experiment 3 found no reliable effect of mediation.